HIV Database Workshop

www.hiv.lanl.gov
seq-info@lanl.gov

Presenters: Bette Korber, Brian Foley & Will FIscher

Database PIs: Bette Korber, Thomas Leitner,
Karina Yusim

Additional database staff: Werner Abfalterer,
Peter Hraber, Elisabeth Sharon Fung, Robert Funkhouser,
Kumkum Ganguly, Jenni Macke, James Szinger,
and Hyejin Yoon

Project Officer: Stuart Shapiro, NIAID, NIH

Theoretical Biology and Biophysics, T-6
Los Alamos National Laboratory

Los Alamos
NATIONAL LABORATORY

Workshop Topics

HIV Sequence Database and Immunology Database

Bette Korber, Brian Foley and Will Fischer

Session 1

Wednesday,
March 12
11:15 – 12:45

General introduction
Sequence search interface – alignments and basic trees
Geography search interface
Histogram
Database Alignments

Tools:

- *Genecutter - processing nucleotide sequences*
- *Neighbor Joining Treemaker*
- *HIV/SIV sequence locator tool*
- *New HIV Gene Map JBROWSE tool*
- *Highlighter*
- *Protein Feature Accent*
- *Quality Control (if time permits)*

Los Alamos
NATIONAL LABORATORY

Workshop Topics

HIV Sequence Database and Immunology Database

Bette Korber, Will Fischer and Brian Foley

Session 2

Thursday,
March 13
11:15 – 12:45

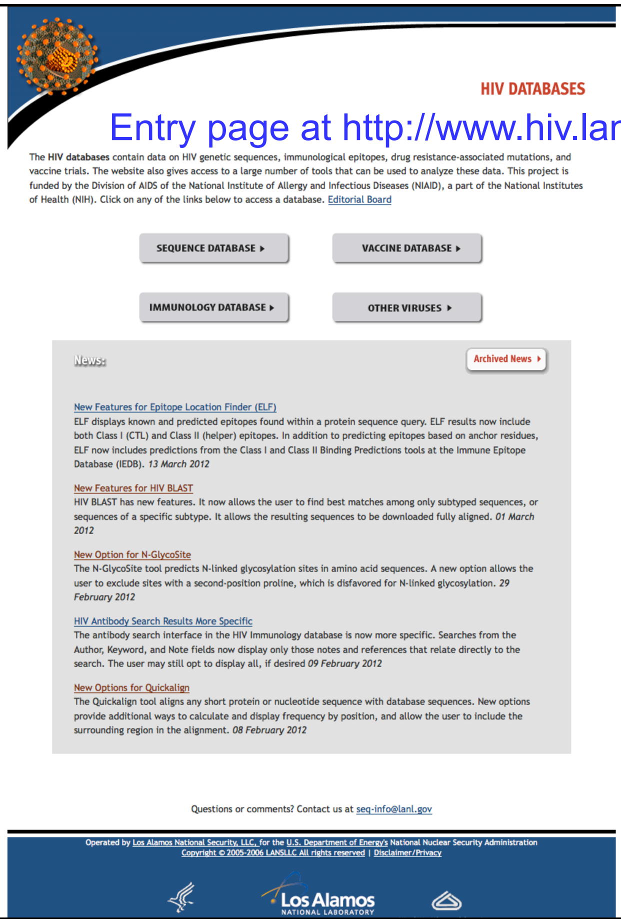
Immunology database introduction
Epitope maps and epitope summary tables
T-cell epitope search
T-cell epitope variants
Antibody search
List of most broadly neutralizing antibodies
HIV/SIV sequence locator tool
QuickAlign – Align an epitope to the database alignments
CATNAP
ELF – epitope location finder
Peptgen – Design peptides for reagent development

Mosaic Vaccine Maker, Epicover, and Posicover
- generate candidate vaccines
- estimate epitope coverage
- determine regional epitope coverage

Workshop Goals

- Understanding the database content, how information was obtained, and what is available
- Database searching
- Examples of tools for analyses
- Quality control tools

Entry page at http://www.hiv.lanl.gov/



The **HIV databases** contain data on HIV genetic sequences, immunological epitopes, drug resistance-associated mutations, and vaccine trials. The website also gives access to a large number of tools that can be used to analyze these data. This project is funded by the Division of AIDS of the National Institute of Allergy and Infectious Diseases (NIAID), a part of the National Institutes of Health (NIH). Click on any of the links below to access a database. Editorial Board

New Features for Epitope Location Finder (ELF)
ELF displays known and predicted epitopes found within a protein sequence query. ELF results now include both Class I (CTL) and Class II (helper) epitopes. In addition to predicting epitopes based on anchor residues, ELF now includes predictions from the Class I and Class II Binding Predictions tools at the Immune Epitope Database (IEDB). *13 March 2012*

New Features for HIV BLAST
HIV BLAST has new features. It now allows the user to find best matches among only subtyped sequences, or sequences of a specific subtype. It allows the resulting sequences to be downloaded fully aligned. *01 March 2012*

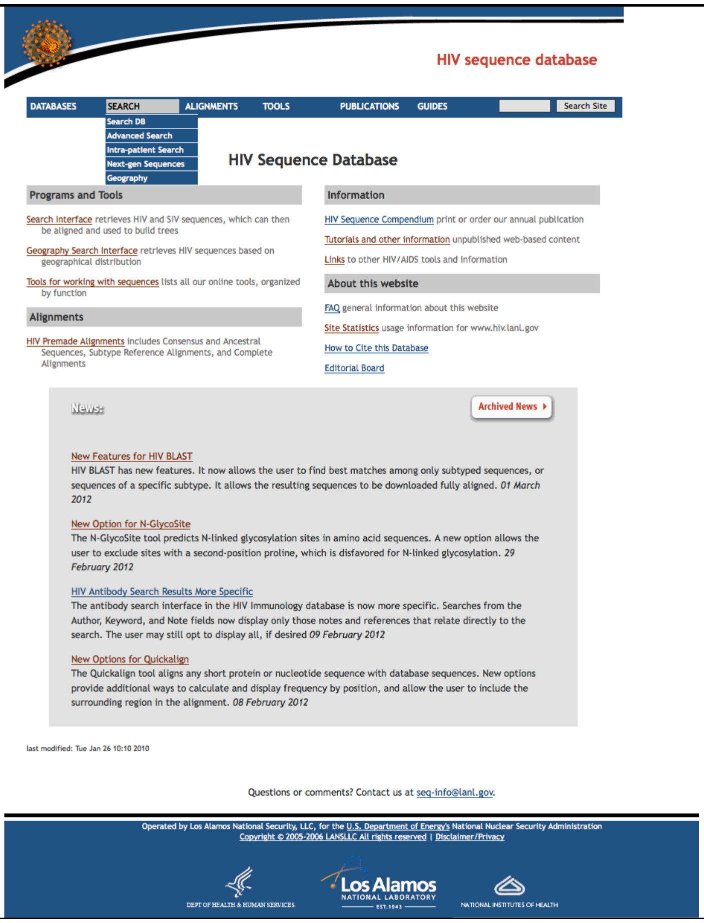
New Option for N-GlycoSite
The N-GlycoSite tool predicts N-linked glycosylation sites in amino acid sequences. A new option allows the user to exclude sites with a second-position proline, which is disfavored for N-linked glycosylation. *29 February 2012*

HIV Antibody Search Results More Specific
The antibody search interface in the HIV Immunology database is now more specific. Searches from the Author, Keyword, and Note fields now display only those notes and references that relate directly to the search. The user may still opt to display all, if desired *09 February 2012*

New Options for Quickalign
The Quickalign tool aligns any short protein or nucleotide sequence with database sequences. New options provide additional ways to calculate and display frequency by position, and allow the user to include the surrounding region in the alignment. *08 February 2012*

Questions or comments? Contact us at seq-info@lanl.gov

Operated by Los Alamos National Security, LLC, for the U.S. Department of Energy's National Nuclear Security Administration
| Disclaimer/Privacy



HIV Sequence Database

Programs and Tools

Search Interface retrieves HIV and SIV sequences, which can then be aligned and used to build trees

Geography Search Interface retrieves HIV sequences based on geographical distribution

Tools for working with sequences lists all our online tools, organized by function

Alignments

HIV Premade Alignments includes Consensus and Ancestral Sequences, Subtype Reference Alignments, and Complete Alignments

Information

HIV Sequence Compendium print or order our annual publication

Tutorials and other information unpublished web-based content

Links to other HIV/AIDS tools and information

About this website

FAQ general information about this website

Site Statistics usage information for www.hiv.lanl.gov

How to Cite this Database

Editorial Board

New Features for HIV BLAST
HIV BLAST has new features. It now allows the user to find best matches among only subtyped sequences, or sequences of a specific subtype. It allows the resulting sequences to be downloaded fully aligned. *01 March 2012*

New Option for N-GlycoSite
The N-GlycoSite tool predicts N-linked glycosylation sites in amino acid sequences. A new option allows the user to exclude sites with a second-position proline, which is disfavored for N-linked glycosylation. *29 February 2012*

HIV Antibody Search Results More Specific
The antibody search interface in the HIV Immunology database is now more specific. Searches from the Author, Keyword, and Note fields now display only those notes and references that relate directly to the search. The user may still opt to display all, if desired *09 February 2012*

New Options for Quickalign
The Quickalign tool aligns any short protein or nucleotide sequence with database sequences. New options provide additional ways to calculate and display frequency by position, and allow the user to include the surrounding region in the alignment. *08 February 2012*

last modified: Tue Jan 26 10:10 2010

Questions or comments? Contact us at seq-info@lanl.gov.

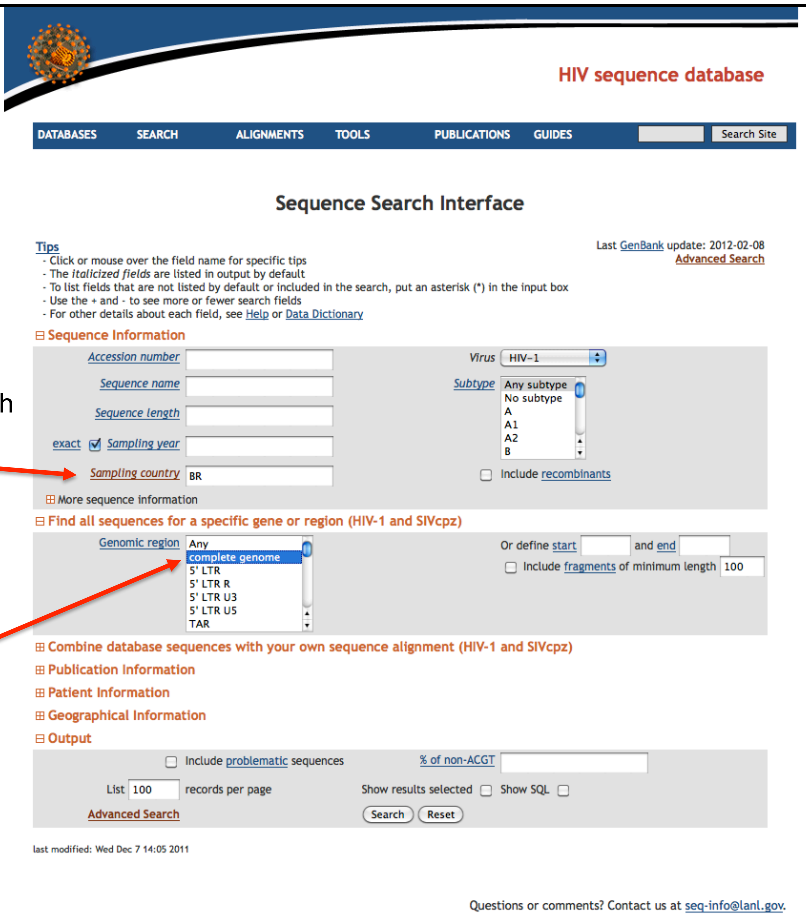
Operated by Los Alamos National Security, LLC, for the U.S. Department of Energy's National Nuclear Security Administration
| Disclaimer/Privacy

Search Interface

- **Help**
  - Tips at the top of the page are often overlooked
    - Ranges, operators, wildcards, logical groupings
  - Mouse-over provides brief descriptions; click field names for details in Help file
- **Searches**
  - Searches are case-insensitive
  - Records are searchable through sequence, patient, genomic region, or publication information and can be matched to the genomic region of a user-provided alignment
  - First seven fields will appear in search results page by default
  - A “*” in a textbox will cause that field to be included in the results page
  - Patient information (Infection year, Infection country) is different than sequence information (Sampling year and Sampling country)
  - Problematic sequence filters (hypermutation, frequent ambiguities, potential contamination)
- **Analysis**
  - Build a tree with user alignment, search results and subtype reference sequences combined
- **Results**
  - Can download aligned or unaligned sequences
  - Alignments are based on multiple pair wise alignments – alignments are good, but need hand editing for an optimal alignment
  - Select all or a subset of sequences for download
  - Sequences can be re-ordered by clicking on fields at the top of the page

We will search for country = Brazil (BR)

We will search for complete genomes.



Sequence Search Interface

Tips

- Click or mouse over the field name for specific tips
- The *italicized fields* are listed in output by default
- To list fields that are not listed by default or included in the search, put an asterisk (*) in the input box
- Use the + and - to see more or fewer search fields
- For other details about each field, see Help or Data Dictionary

Last GenBank update: 2012-02-08
Advanced Search

⊟ Sequence Information

Accession number
Sequence name
Sequence length
exact ☑ *Sampling year*
Sampling country BR
Virus HIV-1
Subtype Any subtype / No subtype / A / A1 / A2 / B
□ Include recombinants
⊞ More sequence information

⊟ Find all sequences for a specific gene or region (HIV-1 and SIVcpz)

Genomic region: Any / **complete genome** / 5' LTR / 5' LTR R / 5' LTR U3 / 5' LTR U5 / TAR
Or define start and end
□ Include fragments of minimum length 100

⊞ Combine database sequences with your own sequence alignment (HIV-1 and SIVcpz)

⊞ Publication Information

⊞ Patient Information

⊞ Geographical Information

⊟ Output

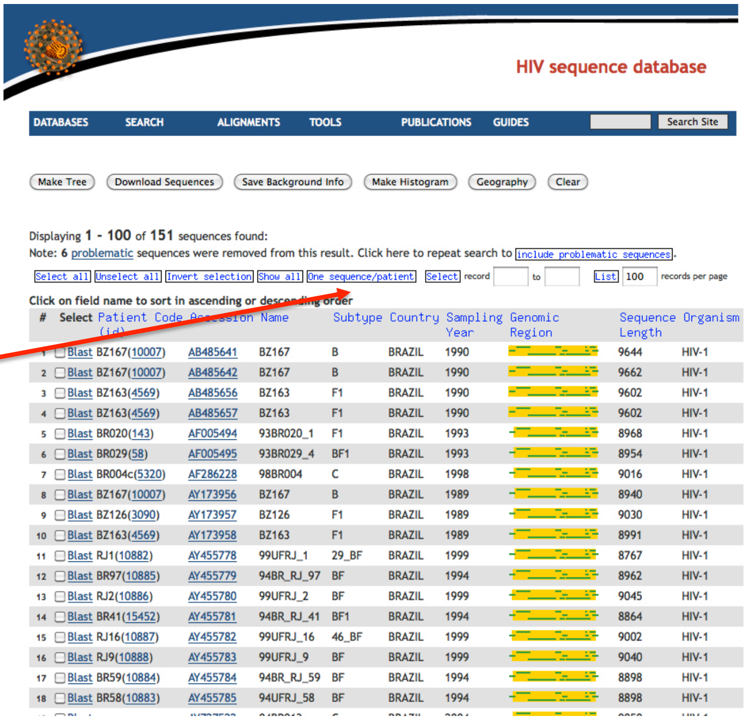
□ Include problematic sequences
% of non-ACGT
List 100 records per page
Show results selected □ Show SQL □
Advanced Search
Search Reset

last modified: Wed Dec 7 14:05 2011

Questions or comments? Contact us at seq-info@lanl.gov.

Results for HIV-1 complete genomes from Brazil

Choose “One sequence/ patient” to remove very similar sequences (only available if a region is selected)

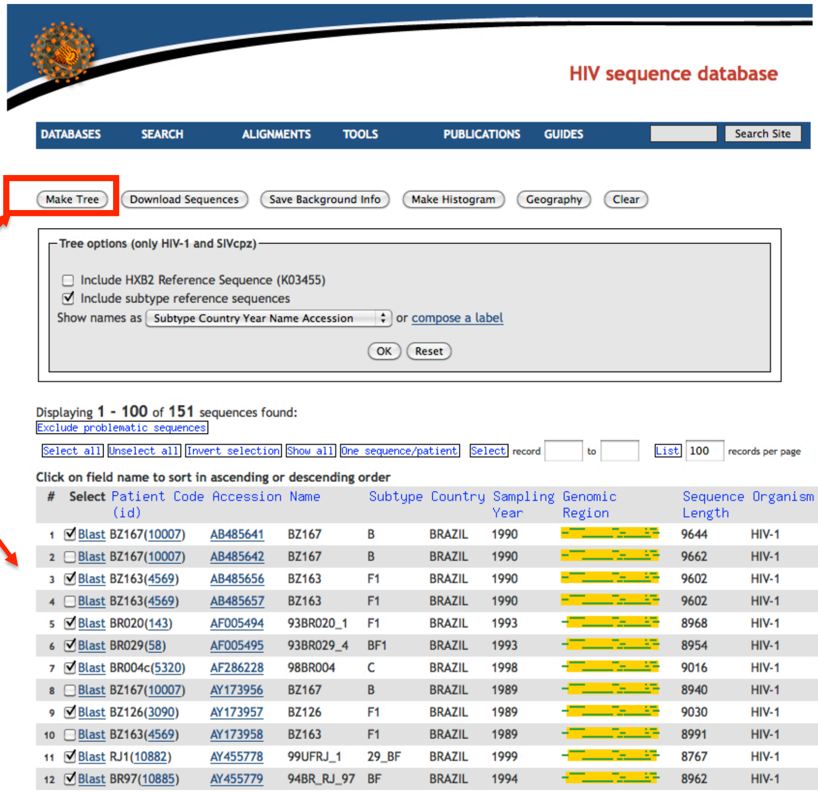


Displaying **1 - 100** of **151** sequences found:
Note: 6 problematic sequences were removed from this result. Click here to repeat search to include problematic sequences.

Click on field name to sort in ascending or descending order

#	Select	Patient Code (id)	Accession	Name	Subtype	Country	Sampling Year	Genomic Region	Sequence Length	Organism
1	□	Blast BZ167(10007)	AB485641	BZ167	B	BRAZIL	1990		9644	HIV-1
2	□	Blast BZ167(10007)	AB485642	BZ167	B	BRAZIL	1990		9662	HIV-1
3	□	Blast BZ163(4569)	AB485656	BZ163	F1	BRAZIL	1990		9602	HIV-1
4	□	Blast BZ163(4569)	AB485657	BZ163	F1	BRAZIL	1990		9602	HIV-1
5	□	Blast BR020(143)	AF005494	93BR020_1	F1	BRAZIL	1993		8968	HIV-1
6	□	Blast BR029(58)	AF005495	93BR029_4	BF1	BRAZIL	1993		8954	HIV-1
7	□	Blast BR004c(5320)	AF286228	98BR004	C	BRAZIL	1998		9016	HIV-1
8	□	Blast BZ167(10007)	AY173956	BZ167	B	BRAZIL	1989		8940	HIV-1
9	□	Blast BZ126(3090)	AY173957	BZ126	F1	BRAZIL	1989		9030	HIV-1
10	□	Blast BZ163(4569)	AY173958	BZ163	F1	BRAZIL	1989		8991	HIV-1
11	□	Blast RJ1(10882)	AY455778	99UFRJ_1	29_BF	BRAZIL	1999		8767	HIV-1
12	□	Blast BR97(10885)	AY455779	94BR_RJ_97	BF	BRAZIL	1994		8962	HIV-1
13	□	Blast RJ2(10886)	AY455780	99UFRJ_2	BF	BRAZIL	1999		9045	HIV-1
14	□	Blast BR41(15452)	AY455781	94BR_RJ_41	BF1	BRAZIL	1994		8864	HIV-1
15	□	Blast RJ16(10887)	AY455782	99UFRJ_16	46_BF	BRAZIL	1999		9002	HIV-1
16	□	Blast RJ9(10888)	AY455783	99UFRJ_9	BF	BRAZIL	1999		9040	HIV-1
17	□	Blast BR59(10884)	AY455784	94BR_RJ_59	BF	BRAZIL	1994		8898	HIV-1
18	□	Blast BR58(10883)	AY455785	94UFRJ_58	BF	BRAZIL	1994		8898	HIV-1

Select a few sequences and make tree, allows us to add a reference set to our data and align them



Tree options (only HIV-1 and SIVcpz)

□ Include HXB2 Reference Sequence (K03455)
☑ Include subtype reference sequences
Show names as: Subtype Country Year Name Accession or compose a label

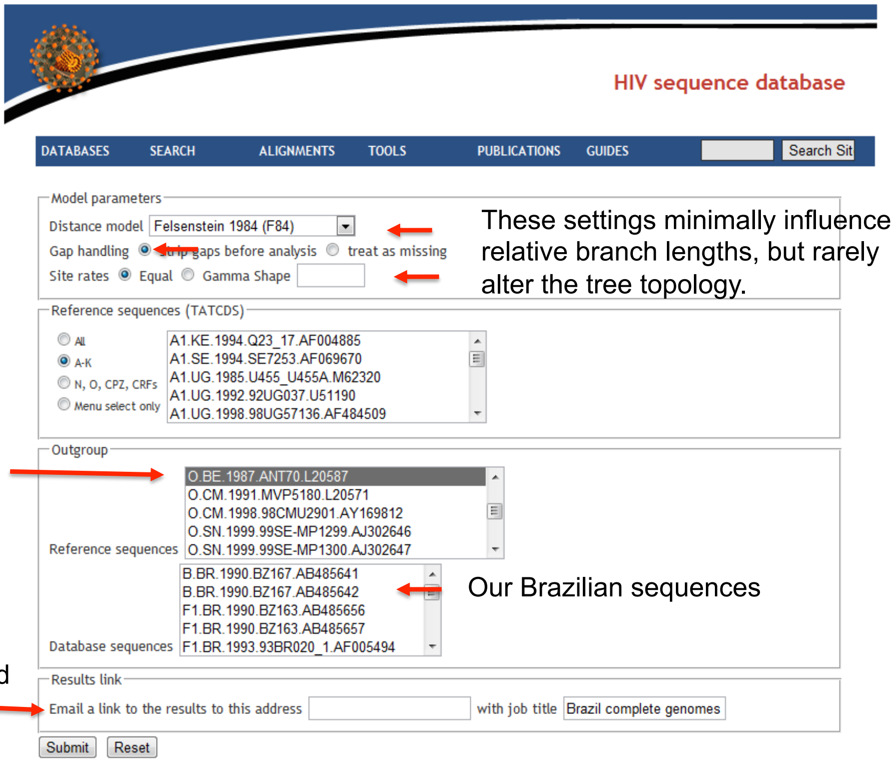
Displaying **1 - 100** of **151** sequences found:

Click on field name to sort in ascending or descending order

#	Select	Patient Code (id)	Accession	Name	Subtype	Country	Sampling Year	Genomic Region	Sequence Length	Organism
1	☑	Blast BZ167(10007)	AB485641	BZ167	B	BRAZIL	1990		9644	HIV-1
2	□	Blast BZ167(10007)	AB485642	BZ167	B	BRAZIL	1990		9662	HIV-1
3	☑	Blast BZ163(4569)	AB485656	BZ163	F1	BRAZIL	1990		9602	HIV-1
4	□	Blast BZ163(4569)	AB485657	BZ163	F1	BRAZIL	1990		9602	HIV-1
5	☑	Blast BR020(143)	AF005494	93BR020_1	F1	BRAZIL	1993		8968	HIV-1
6	☑	Blast BR029(58)	AF005495	93BR029_4	BF1	BRAZIL	1993		8954	HIV-1
7	☑	Blast BR004c(5320)	AF286228	98BR004	C	BRAZIL	1998		9016	HIV-1
8	□	Blast BZ167(10007)	AY173956	BZ167	B	BRAZIL	1989		8940	HIV-1
9	☑	Blast BZ126(3090)	AY173957	BZ126	F1	BRAZIL	1989		9030	HIV-1
10	□	Blast BZ163(4569)	AY173958	BZ163	F1	BRAZIL	1989		8991	HIV-1
11	☑	Blast RJ1(10882)	AY455778	99UFRJ_1	29_BF	BRAZIL	1999		8767	HIV-1
12	☑	Blast BR97(10885)	AY455779	94BR_RJ_97	BF	BRAZIL	1994		8962	HIV-1

TreeMaker tool

Choice of outgroup influences the the tree. In general, choose next closest sequences to the “ingroup”. In this case our Brazilian sequences are all HIV-1 M group.



Model parameters

Distance model: Felsenstein 1984 (F84)

Gap handling ● strip gaps before analysis ○ treat as missing

Site rates ● Equal ○ Gamma Shape

These settings minimally influence relative branch lengths, but rarely alter the tree topology.

Reference sequences (TATCDS)

○ All
● A-K
○ N, O, CPZ, CRFs
○ Menu select only

A1.KE.1994.Q23_17.AF004885
A1.SE.1994.SE7253.AF069670
A1.UG.1985.U455_U455A.M62320
A1.UG.1992.92UG037.U51190
A1.UG.1998.98UG57136.AF484509

Outgroup

Reference sequences

O.BE.1987.ANT70.L20587
O.CM.1991.MVP5180.L20571
O.CM.1998.98CMU2901.AY169812
O.SN.1999.99SE-MP1299.AJ302646
O.SN.1999.99SE-MP1300.AJ302647

Database sequences

B.BR.1990.BZ167.AB485641
B.BR.1990.BZ167.AB485642
F1.BR.1990.BZ163.AB485656
F1.BR.1990.BZ163.AB485657
F1.BR.1993.93BR020_1.AF005494

Our Brazilian sequences

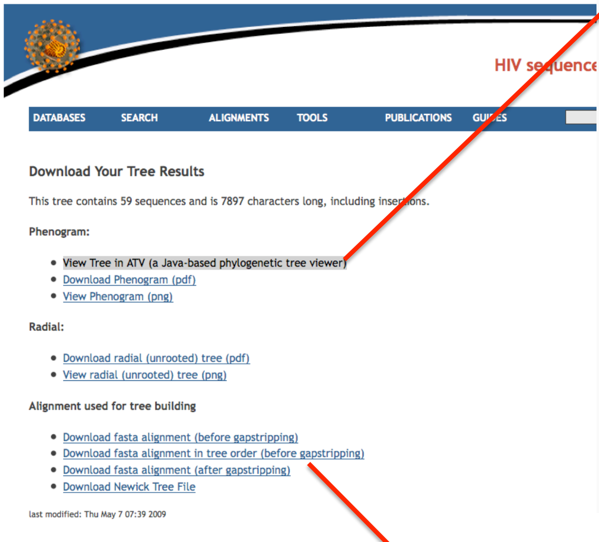
Results link

Email a link to the results to this address [ ] with job title Brazil complete genomes

Submit Reset

Optional mailback, and tree title

ATV java-based view for quick look, cannot save/print



Download Your Tree Results

This tree contains 59 sequences and is 7897 characters long, including insertions.

Phenogram:

- View Tree in ATV (a Java-based phylogenetic tree viewer)
- Download Phenogram (pdf)
- View Phenogram (png)

Radial:

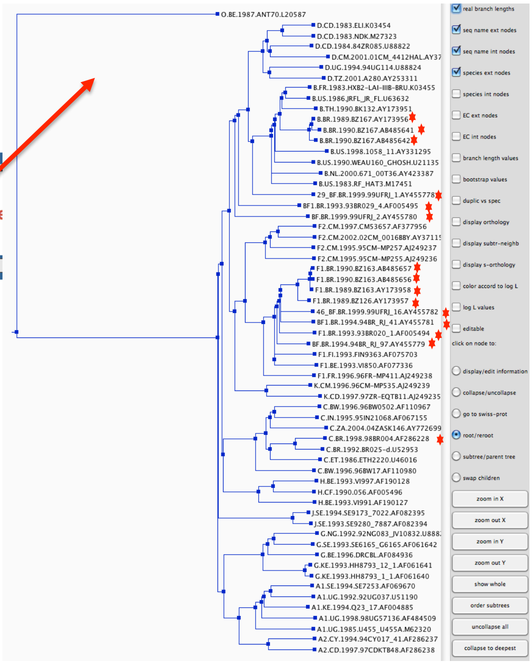
- Download radial (unrooted) tree (pdf)
- View radial (unrooted) tree (png)

Alignment used for tree building

- Download fasta alignment (before gapstripping)
- Download fasta alignment in tree order (before gapstripping)
- Download fasta alignment (after gapstripping)
- Download Newick Tree File

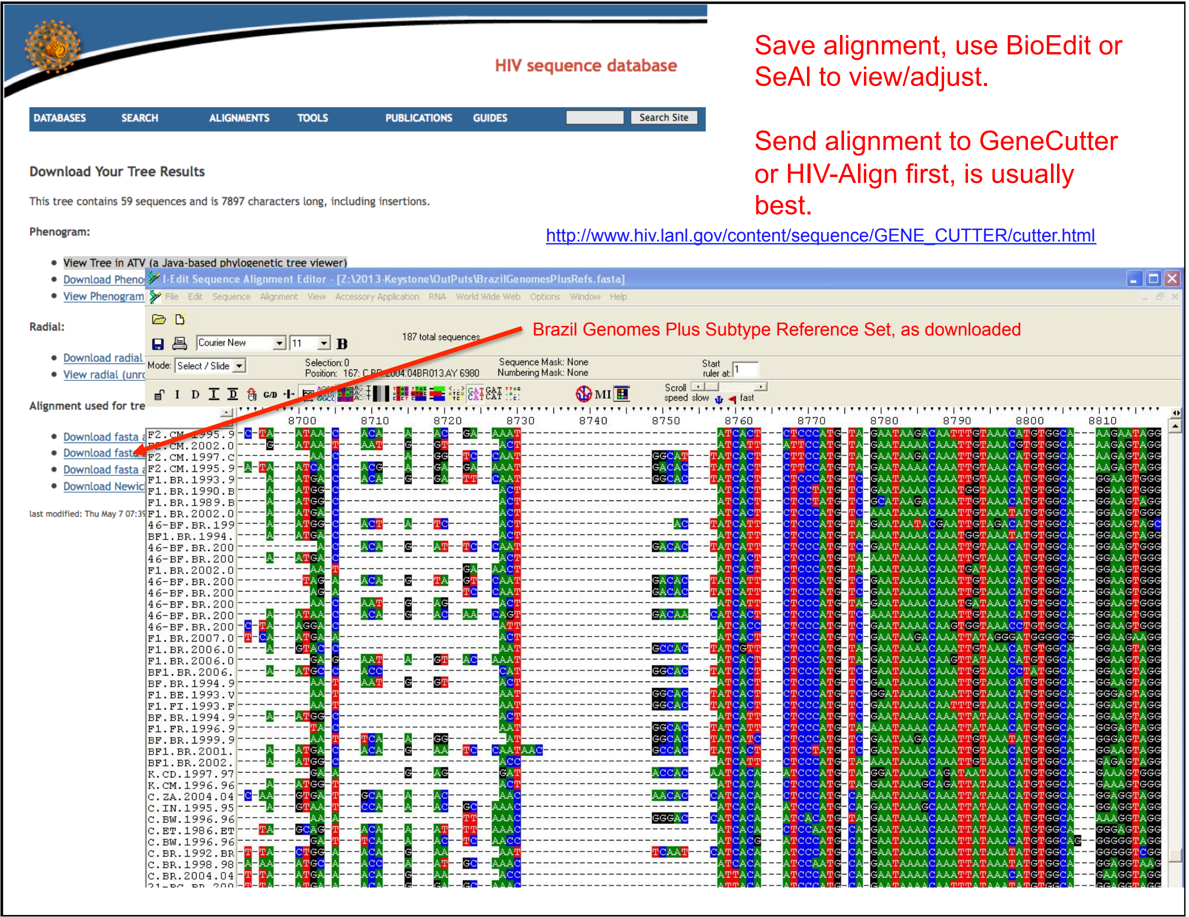
last modified: Thu May 7 07:39 2009

Questions or comments? Contact us at seq-info@lanl.gov.


O.BE.1987.ANT70.L20587
D.CD.1983.ELI.K03454
D.CD.1983.NDK.M27323
D.CD.1984.84ZR085.U88822
D.CM.2001.01CM_4412HAL.AY37
D.UG.1994.94UG114.U88824
D.TZ.2001.A280.AY253311
B.FR.1983.HXB2-LAI-IIIB-BRU.K03455
B.US.1986.JRFL_JR_FL.U63632
B.TH.1990.BK132.AY173951
B.BR.1989.BZ167.AY173956
B.BR.1990.BZ167.AB485641
B.BR.1990.BZ167.AB485642
B.US.1998.1058_11.AY331295
B.US.1990.WEAU160_GHOSH.U21135
B.NL.2000.671_00T36.AY423387
B.US.1983.RF_HAT3.M17451
29_BF.BR.1999.99UFRJ_1.AY455778
BF1.BR.1993.93BR029_4.AF005495
BF.BR.1999.99UFRJ_2.AY455780
F2.CM.1997.CM53657.AF377956
F2.CM.2002.02CM_0016BBY.AY3711
F2.CM.1995.95CM-MP257.AJ249237
F2.CM.1995.95CM-MP255.AJ249236
F1.BR.1990.BZ163.AB485657
F1.BR.1990.BZ163.AB485656
F1.BR.1989.BZ163.AY173958
F1.BR.1989.BZ126.AY173957
46_BF.BR.1999.99UFRJ_16.AY455782
BF1.BR.1994.94BR_RJ_41.AY455781
F1.BR.1993.93BR020_1.AF005494
BF.BR.1994.94BR_RJ_97.AY455779
F1.FI.1993.FIN9363.AF075703
F1.BE.1993.VI850.AF077336
F1.FR.1996.96FR-MP411.AJ249238
K.CM.1996.96CM-MP535.AJ249239
K.CD.1997.97ZR-EQTB11.AJ249235
C.BW.1996.96BW0502.AF110967
C.IN.1995.95IN21068.AF067155
C.ZA.2004.04ZASK146.AY772699
C.BR.1998.98BR004.AF286228
C.BR.1992.BR025-d.U52953
C.ET.1986.ETH2220.U46016
C.BW.1996.96BW17.AF110980
H.BE.1993.VI997.AF190128
H.CF.1990.056.AF005496
H.BE.1993.VI991.AF190127
J.SE.1994.SE9173_7022.AF082395
J.SE.1993.SE9280_7887.AF082394
G.NG.1992.92NG083_JV10832.U888
G.SE.1993.SE6165_G6165.AF061642
G.BE.1996.DRCBL.AF084936
G.KE.1993.HH8793_12_1.AF061641
G.KE.1993.HH8793_1_1.AF061640
A1.SE.1994.SE7253.AF069670
A1.UG.1992.92UG037.U51190
A1.KE.1994.Q23_17.AF004885
A1.UG.1998.98UG57136.AF484509
A1.UG.1985.U455_U455A.M62320
A2.CY.1994.94CY017_41.AF286237
A2.CD.1997.97CDKTB48.AF286238
☑ real branch lengths
☑ seq name ext nodes
☑ seq name int nodes
☑ species ext nodes
☐ species int nodes
☐ EC ext nodes
☐ EC int nodes
☐ branch length values
☐ bootstrap values
☐ duplic vs spec
☐ display orthology
☐ display subtr-neighb
☐ display s-orthology
☐ color accord to log L
☐ log L values
☐ editable
click on node to:
○ display/edit information
○ collapse/uncollapse
○ go to swiss-prot
● root/reroot
○ subtree/parent tree
○ swap children
zoom in X
zoom out X
zoom in Y
zoom out Y
show whole
order subtrees
uncollapse all
collapse to deepest

Obtaining your sequences of interest and having them aligned to a good reference set was the whole point of this. The tree was just a first check on data and alignment quality.

Save alignment, use BioEdit or SeAl to view/adjust.



Save alignment, use BioEdit or SeAl to view/adjust.

Send alignment to GeneCutter or HIV-Align first, is usually best.

http://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html

Brazil Genomes Plus Subtype Reference Set, as downloaded

Download Your Tree Results

This tree contains 59 sequences and is 7897 characters long, including insertions.

Phenogram:

- View Tree in ATV (a Java-based phylogenetic tree viewer)
- Download Pheno
- View Phenogram

Radial:

- Download radial
- View radial (unr

Alignment used for tre

- Download fasta a
- Download fasta
- Download fasta a
- Download Newic

last modified: Thu May 7 07:3

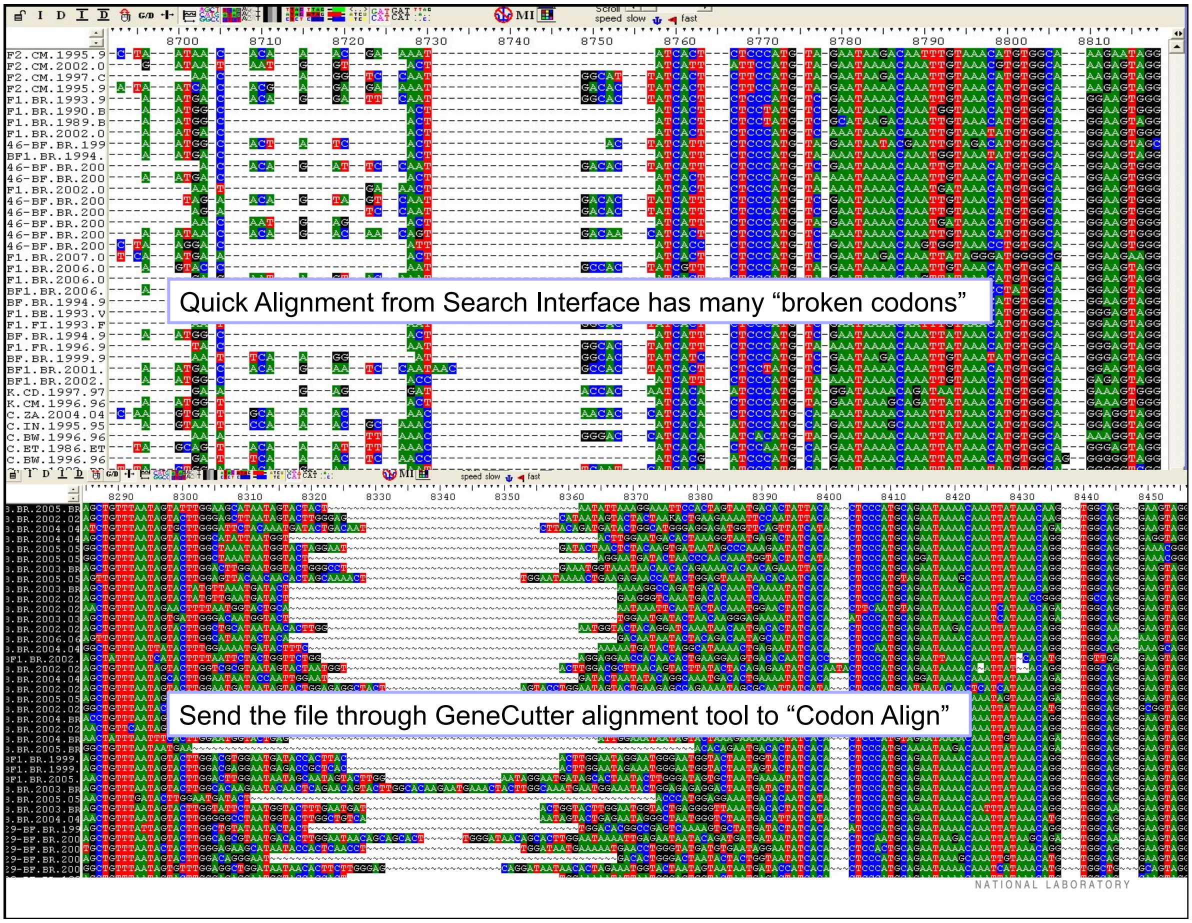
I-Edit Sequence Alignment Editor - [Z:\2013-Keystone\OutPuts\BrazilGenomesPlusRefs.fasta]

187 total sequences

Mode: Select / Slide

Selection: 0
Position: 167: C.BR.2004.04BR013.AY 6980

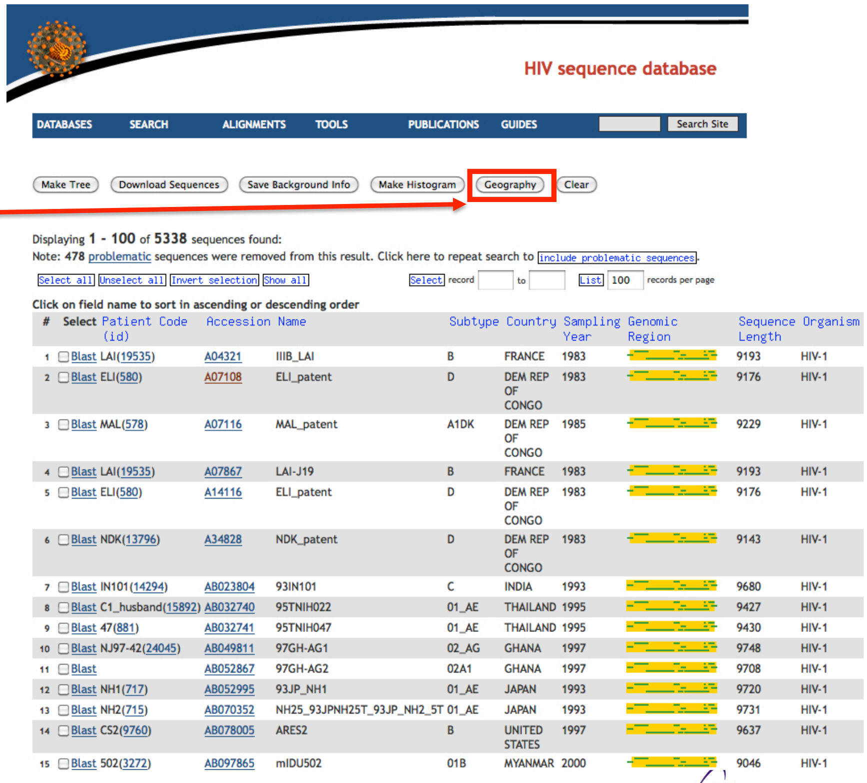
Sequence Mask: None
Numbering Mask: None



Quick Alignment from Search Interface has many “broken codons”

Send the file through GeneCutter alignment tool to “Codon Align”

New search: all complete genomes; then look at geographic and subtype distribution of the sequences

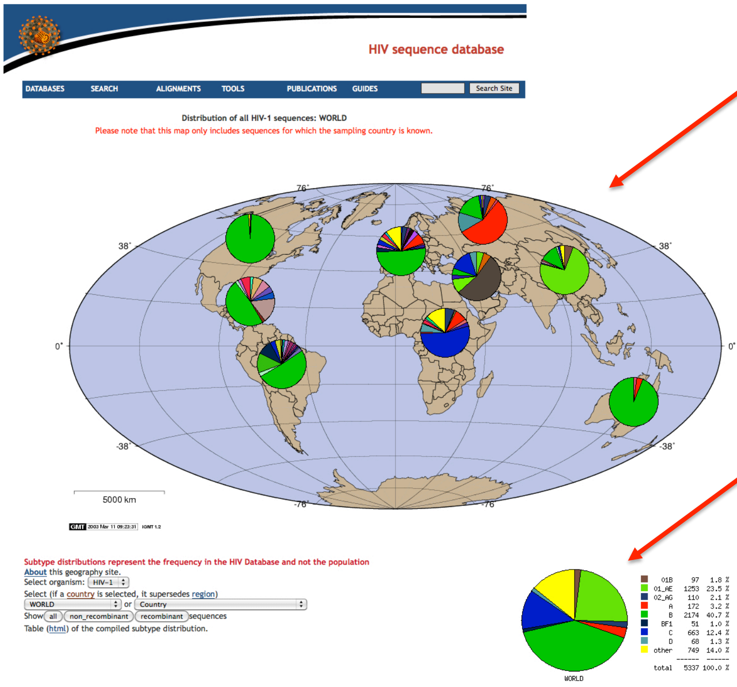


Displaying **1 - 100** of **5338** sequences found:
Note: 478 problematic sequences were removed from this result. Click here to repeat search to include problematic sequences.

Click on field name to sort in ascending or descending order

#	Select	Patient Code (id)	Accession	Name	Subtype	Country	Sampling Year	Genomic Region	Sequence Length	Organism
1	□ Blast	LAI(19535)	A04321	IIIB_LAI	B	FRANCE	1983		9193	HIV-1
2	□ Blast	ELI(580)	A07108	ELI_patent	D	DEM REP OF CONGO	1983		9176	HIV-1
3	□ Blast	MAL(578)	A07116	MAL_patent	A1DK	DEM REP OF CONGO	1985		9229	HIV-1
4	□ Blast	LAI(19535)	A07867	LAI-J19	B	FRANCE	1983		9193	HIV-1
5	□ Blast	ELI(580)	A14116	ELI_patent	D	DEM REP OF CONGO	1983		9176	HIV-1
6	□ Blast	NDK(13796)	A34828	NDK_patent	D	DEM REP OF CONGO	1983		9143	HIV-1
7	□ Blast	IN101(14294)	AB023804	93IN101	C	INDIA	1993		9680	HIV-1
8	□ Blast	C1_husband(15892)	AB032740	95TNIH022	01_AE	THAILAND	1995		9427	HIV-1
9	□ Blast	47(881)	AB032741	95TNIH047	01_AE	THAILAND	1995		9430	HIV-1
10	□ Blast	NJ97-42(24045)	AB049811	97GH-AG1	02_AG	GHANA	1997		9748	HIV-1
11	□ Blast		AB052867	97GH-AG2	02A1	GHANA	1997		9708	HIV-1
12	□ Blast	NH1(717)	AB052995	93JP_NH1	01_AE	JAPAN	1993		9720	HIV-1
13	□ Blast	NH2(715)	AB070352	NH25_93JPNH25T_93JP_NH2_5T	01_AE	JAPAN	1993		9731	HIV-1
14	□ Blast	CS2(9760)	AB078005	ARES2	B	UNITED STATES	1997		9637	HIV-1
15	□ Blast	502(3272)	AB097865	mIDU502	01B	MYANMAR	2000		9046	HIV-1

Geography output



Distribution of all HIV-1 sequences: WORLD
Please note that this map only includes sequences for which the sampling country is known.

Subtype distributions represent the frequency in the HIV Database and not the population
About this geography site.
Select organism: HIV-1
Select (if a country is selected, it supersedes region)
WORLD or Country
Show all non_recombinant recombinant sequences
Table (html) of the compiled subtype distribution.

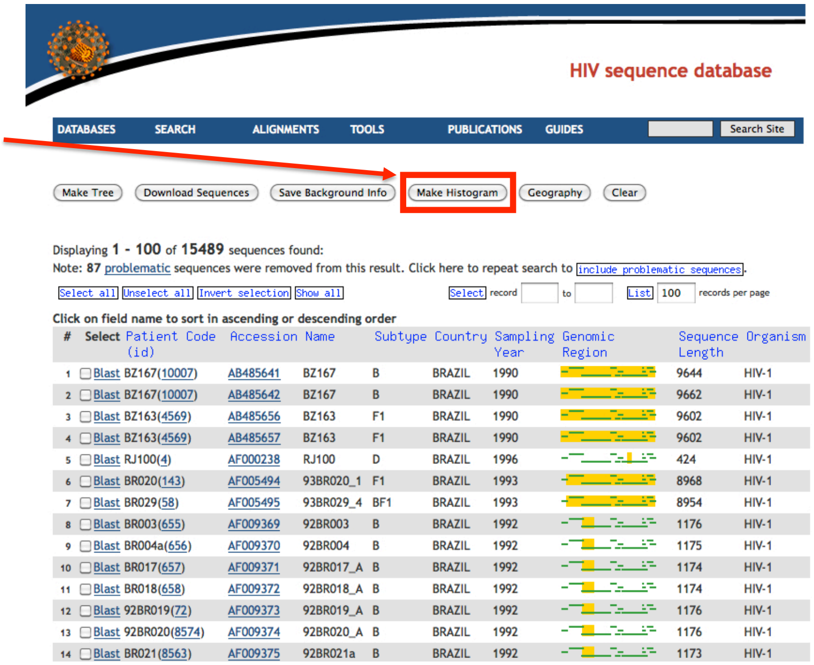
```
01B      97    1.8 %
01_AE  1253   23.5 %
02_AG   110    2.1 %
A       172    3.2 %
B      2174   40.7 %
BF1      51    1.0 %
C       663   12.4 %
D        68    1.3 %
other   749   14.0 %
       ------ ------
total  5337  100.0 %
```

Each continent's pie chart is clickable to "zoom in" on that continent.

Likewise for each country once you are zoomed in to the continent level.

Most complete genomes in the HIV database are subtype B. But subtype C is more prevalent in human infections. Beware of this type of sampling bias.

New search: all sequences from Brazil. Then look at the distribution of the sequences over the genome



HIV sequence database

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS GUIDES Search Site

Make Tree | Download Sequences | Save Background Info | Make Histogram | Geography | Clear

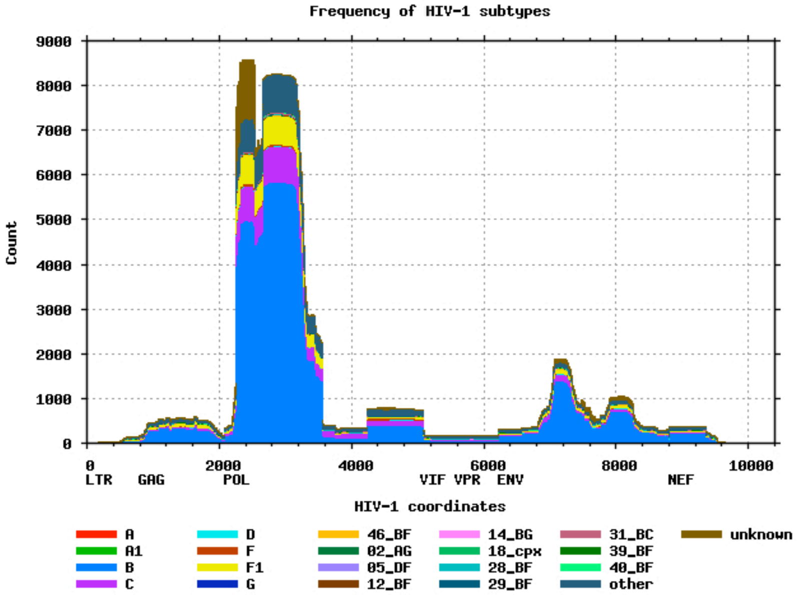
Displaying **1 - 100** of **15489** sequences found:
Note: **87** problematic sequences were removed from this result. Click here to repeat search to include problematic sequences.

Select all | Unselect all | Invert selection | Show all   Select record __ to __   List 100 records per page

Click on field name to sort in ascending or descending order

#	Select	Patient Code (id)	Accession	Name	Subtype	Country	Sampling Year	Genomic Region	Sequence Length	Organism
1	☐ Blast	BZ167(10007)	AB485641	BZ167	B	BRAZIL	1990		9644	HIV-1
2	☐ Blast	BZ167(10007)	AB485642	BZ167	B	BRAZIL	1990		9662	HIV-1
3	☐ Blast	BZ163(4569)	AB485656	BZ163	F1	BRAZIL	1990		9602	HIV-1
4	☐ Blast	BZ163(4569)	AB485657	BZ163	F1	BRAZIL	1990		9602	HIV-1
5	☐ Blast	RJ100(4)	AF000238	RJ100	D	BRAZIL	1996		424	HIV-1
6	☐ Blast	BR020(143)	AF005494	93BR020_1	F1	BRAZIL	1993		8968	HIV-1
7	☐ Blast	BR029(58)	AF005495	93BR029_4	BF1	BRAZIL	1993		8954	HIV-1
8	☐ Blast	BR003(655)	AF009369	92BR003	B	BRAZIL	1992		1176	HIV-1
9	☐ Blast	BR004a(656)	AF009370	92BR004	B	BRAZIL	1992		1175	HIV-1
10	☐ Blast	BR017(657)	AF009371	92BR017_A	B	BRAZIL	1992		1174	HIV-1
11	☐ Blast	BR018(658)	AF009372	92BR018_A	B	BRAZIL	1992		1174	HIV-1
12	☐ Blast	92BR019(72)	AF009373	92BR019_A	B	BRAZIL	1992		1176	HIV-1
13	☐ Blast	92BR020(8574)	AF009374	92BR020_A	B	BRAZIL	1992		1176	HIV-1
14	☐ Blast	BR021(8563)	AF009375	92BR021a	B	BRAZIL	1992		1173	HIV-1

Histogram output



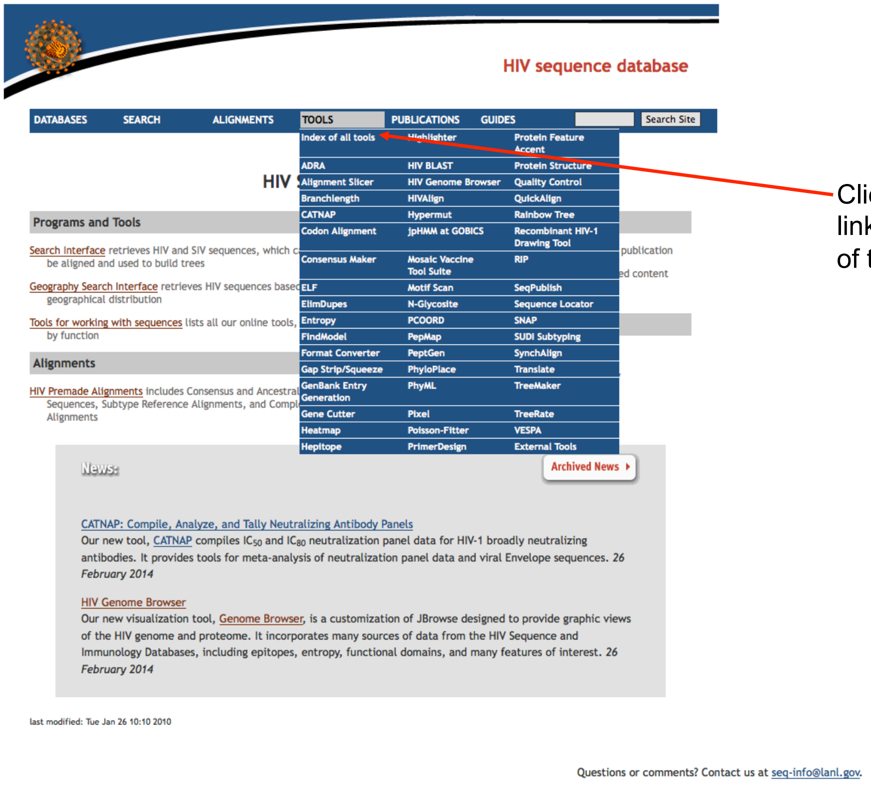
This histogram shows the distribution of sequences from your query across the entire HIV-1 genome. At each position across the genome, the number of sequences overlapping with that position is plotted. The colors represent different subtypes.

Tools

- **Analysis and Quality Control**
  - □ **HIV BLAST** finds sequences similar to yours in the HIV database.
  - □ **N-Glycosite** finds potential N-linked glycosylation sites.
  - □ **RIP 3.0** (Recombinant Identification Program) detects HIV-1 subtypes and recombination.
- **Alignment and sequence manipulation**
  - □ **HIValign** uses our HMM alignment models to align your sequences.
  - □ **Gapstreeze** removes columns with more than a given % of gaps.
  - □ **EpimDupes** Given an alignment or set of unaligned nucleotide or protein sequences, this tool compares the sequences and eliminates any duplicates.
- **Phylogenetics**
  - □ **TreeMaker** generates a neighbor-joining phylogenetic tree.
  - □ **PhyML** generates a maximum likelihood phylogenetic tree.
  - □ **TreeRate** finds the phylogenetic root of a tree and calculates evolutionary rate.
- **Format and display**
  - □ **Protein Feature Accent** provides an interactive 3-D graphic of HIV proteins; the user can map a sequence feature (a short functional domain, epitope, or amino acid) and see where it occurs spatially in the 3D structure.
  - □ **Highlighter** highlights mismatches, matches, transition and transversion mutations, and silent and non-silent mutations in an alignment of nucleotide sequences.
  - □ **SeqPublish** makes alignment publication-ready.
  - □ **Recombinant HIV drawing tool** highlights regions of the genome on a graphically representation

Los Alamos
NATIONAL LABORATORY

The HIV database sequence analysis tool set



Index of all tools	Highlighter	Protein Feature Accent
ADRA	HIV BLAST	Protein Structure
Alignment Slicer	HIV Genome Browser	Quality Control
Branchlength	HIVAlign	QuickAlign
CATNAP	Hypermut	Rainbow Tree
Codon Alignment	jpHMM at GOBICS	Recombinant HIV-1 Drawing Tool
Consensus Maker	Mosaic Vaccine Tool Suite	RIP
ELF	Motif Scan	SeqPublish
ElimDupes	N-Glycosite	Sequence Locator
Entropy	PCOORD	SNAP
FindModel	PepMap	SUDI Subtyping
Format Converter	PeptGen	SynchAlign
Gap Strip/Squeeze	PhyloPlace	Translate
GenBank Entry Generation	PhyML	TreeMaker
Gene Cutter	Pixel	TreeRate
Heatmap	Poisson-Fitter	VESPA
Hepitope	PrimerDesign	External Tools

Programs and Tools

Search Interface retrieves HIV and SIV sequences, which c … publication … be aligned and used to build trees

… ed content

Geography Search Interface retrieves HIV sequences based … geographical distribution

Tools for working with sequences lists all our online tools, … by function

Alignments

HIV Premade Alignments includes Consensus and Ancestral … Sequences, Subtype Reference Alignments, and Compl … Alignments

News:

Archived News ▸

CATNAP: Compile, Analyze, and Tally Neutralizing Antibody Panels
Our new tool, CATNAP compiles IC_{50} and IC_{80} neutralization panel data for HIV-1 broadly neutralizing antibodies. It provides tools for meta-analysis of neutralization panel data and viral Envelope sequences. *26 February 2014*

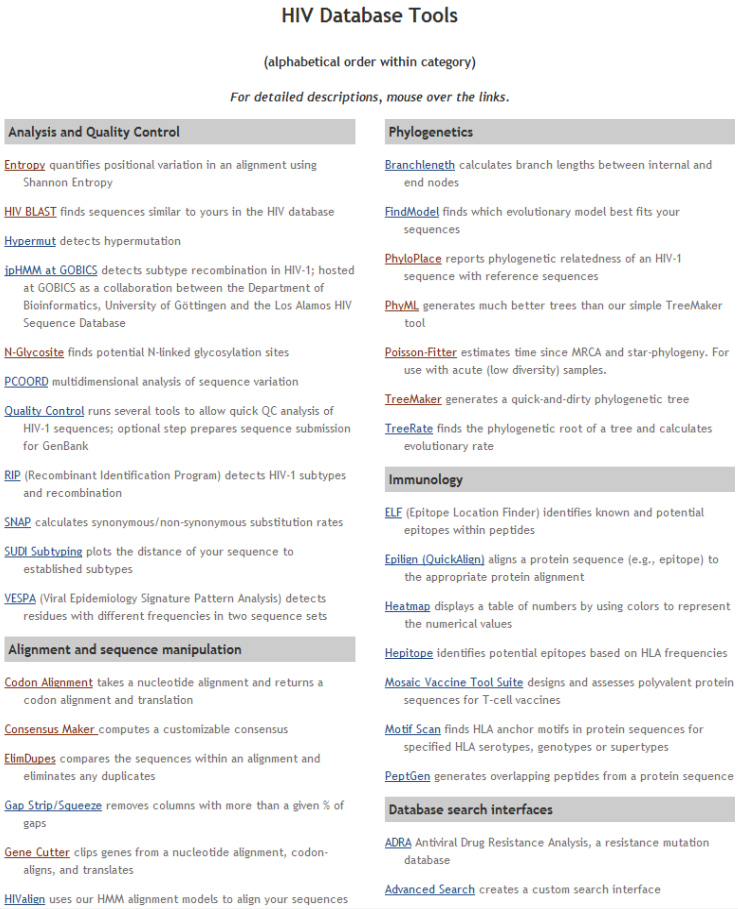
HIV Genome Browser
Our new visualization tool, Genome Browser, is a customization of JBrowse designed to provide graphic views of the HIV genome and proteome. It incorporates many sources of data from the HIV Sequence and Immunology Databases, including epitopes, entropy, functional domains, and many features of interest. *26 February 2014*

last modified: Tue Jan 26 10:10 2010

Questions or comments? Contact us at seq-info@lanl.gov.

Click top level to link to full page of tools

Los Alamos
NATIONAL LABORATORY



HIV Database Tools

(alphabetical order within category)

For detailed descriptions, mouse over the links.

Analysis and Quality Control

Entropy quantifies positional variation in an alignment using Shannon Entropy

HIV BLAST finds sequences similar to yours in the HIV database

Hypermut detects hypermutation

jpHMM at GOBICS detects subtype recombination in HIV-1; hosted at GOBICS as a collaboration between the Department of Bioinformatics, University of Göttingen and the Los Alamos HIV Sequence Database

N-Glycosite finds potential N-linked glycosylation sites

PCOORD multidimensional analysis of sequence variation

Quality Control runs several tools to allow quick QC analysis of HIV-1 sequences; optional step prepares sequence submission for GenBank

RIP (Recombinant Identification Program) detects HIV-1 subtypes and recombination

SNAP calculates synonymous/non-synonymous substitution rates

SUDI Subtyping plots the distance of your sequence to established subtypes

VESPA (Viral Epidemiology Signature Pattern Analysis) detects residues with different frequencies in two sequence sets

Alignment and sequence manipulation

Codon Alignment takes a nucleotide alignment and returns a codon alignment and translation

Consensus Maker computes a customizable consensus

ElimDupes compares the sequences within an alignment and eliminates any duplicates

Gap Strip/Squeeze removes columns with more than a given % of gaps

Gene Cutter clips genes from a nucleotide alignment, codon-aligns, and translates

HIValign uses our HMM alignment models to align your sequences

Phylogenetics

Branchlength calculates branch lengths between internal and end nodes

FindModel finds which evolutionary model best fits your sequences

PhyloPlace reports phylogenetic relatedness of an HIV-1 sequence with reference sequences

PhyML generates much better trees than our simple TreeMaker tool

Poisson-Fitter estimates time since MRCA and star-phylogeny. For use with acute (low diversity) samples.

TreeMaker generates a quick-and-dirty phylogenetic tree

TreeRate finds the phylogenetic root of a tree and calculates evolutionary rate

Immunology

ELF (Epitope Location Finder) identifies known and potential epitopes within peptides

Epilign (QuickAlign) aligns a protein sequence (e.g., epitope) to the appropriate protein alignment

Heatmap displays a table of numbers by using colors to represent the numerical values

Hepitope identifies potential epitopes based on HLA frequencies

Mosaic Vaccine Tool Suite designs and assesses polyvalent protein sequences for T-cell vaccines

Motif Scan finds HLA anchor motifs in protein sequences for specified HLA serotypes, genotypes or supertypes

PeptGen generates overlapping peptides from a protein sequence

Database search interfaces

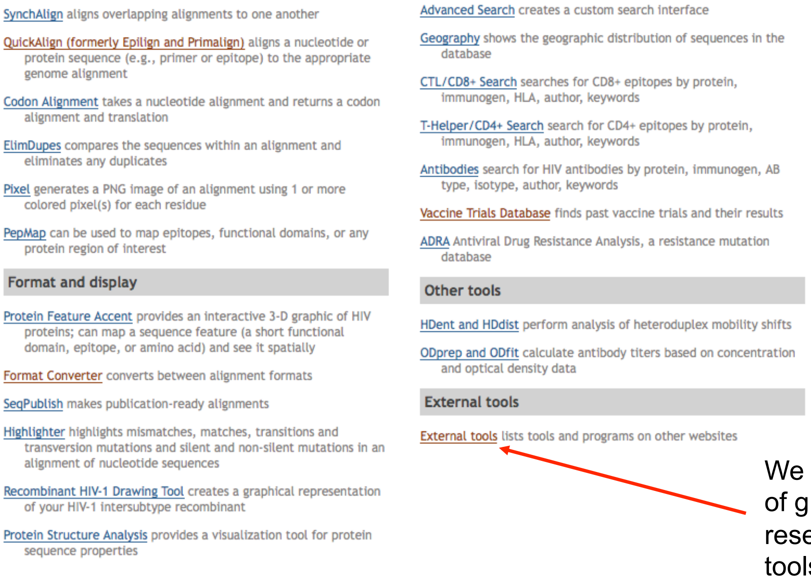
ADRA Antiviral Drug Resistance Analysis, a resistance mutation database

Advanced Search creates a custom search interface

Tools are organized in groups by function/purpose.

Most tools have explanation pages, and sample data sets.

Many tools were inspired by user comments, please ask for more.



SynchAlign aligns overlapping alignments to one another

QuickAlign (formerly Epilign and Primalign) aligns a nucleotide or protein sequence (e.g., primer or epitope) to the appropriate genome alignment

Codon Alignment takes a nucleotide alignment and returns a codon alignment and translation

ElimDupes compares the sequences within an alignment and eliminates any duplicates

Pixel generates a PNG image of an alignment using 1 or more colored pixel(s) for each residue

PepMap can be used to map epitopes, functional domains, or any protein region of interest

Format and display

Protein Feature Accent provides an interactive 3-D graphic of HIV proteins; can map a sequence feature (a short functional domain, epitope, or amino acid) and see it spatially

Format Converter converts between alignment formats

SeqPublish makes publication-ready alignments

Highlighter highlights mismatches, matches, transitions and transversion mutations and silent and non-silent mutations in an alignment of nucleotide sequences

Recombinant HIV-1 Drawing Tool creates a graphical representation of your HIV-1 intersubtype recombinant

Protein Structure Analysis provides a visualization tool for protein sequence properties

Advanced Search creates a custom search interface

Geography shows the geographic distribution of sequences in the database

CTL/CD8+ Search searches for CD8+ epitopes by protein, immunogen, HLA, author, keywords

T-Helper/CD4+ Search search for CD4+ epitopes by protein, immunogen, HLA, author, keywords

Antibodies search for HIV antibodies by protein, immunogen, AB type, isotype, author, keywords

Vaccine Trials Database finds past vaccine trials and their results

ADRA Antiviral Drug Resistance Analysis, a resistance mutation database

Other tools

HDent and HDdist perform analysis of heteroduplex mobility shifts

ODprep and ODfit calculate antibody titers based on concentration and optical density data

External tools

External tools lists tools and programs on other websites

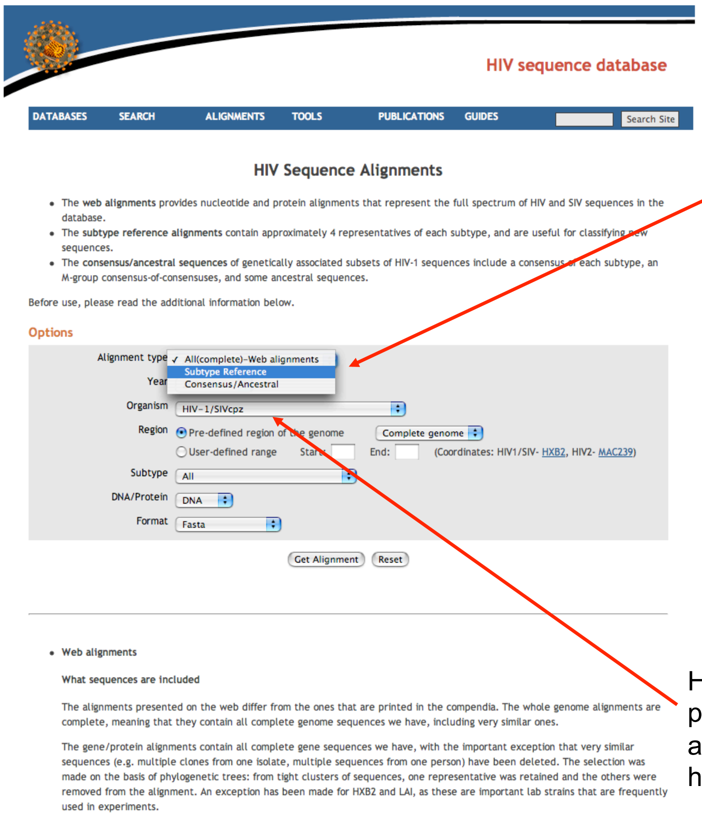
We tend to list only tools of great use in HIV research. Many of these tools are essential, such as either BioEdit or SeAl for alignment viewing and correction.

http://www.hiv.lanl.gov/content/sequence/HIV/HIVTools.html

Pre-Built Sequence alignments

- Originally based on iterations of manual and HMM alignments
- Yearly updates using HMM and manual corrections
- Alignments are in reading frame (codon aligned)
- Contain non-redundant data (one sequence per patient)
- Compendium alignments show fewer sequences than web version
- Reference alignments contain up to four representatives of each subtype. One of each CRF.
- Protein alignments may contain frameshift compensations
- Subtype consensus with ties resolved, as well as maximum likelihood ancestors, are available for reagent production
- Special interest alignments are being added
  - □ Sequence sets of particular research interest
  - □ Suggestions welcome to tkl@lanl.gov

Los Alamos NATIONAL LABORATORY



HIV Sequence Alignments

- The **web alignments** provides nucleotide and protein alignments that represent the full spectrum of HIV and SIV sequences in the database.
- The **subtype reference alignments** contain approximately 4 representatives of each subtype, and are useful for classifying new sequences.
- The **consensus/ancestral sequences** of genetically associated subsets of HIV-1 sequences include a consensus of each subtype, an M-group consensus-of-consensuses, and some ancestral sequences.

Before use, please read the additional information below.

Options

Alignment type: All(complete)-Web alignments
Year
Organism: HIV-1/SIVcpz
Region: ● Pre-defined region of the genome: Complete genome
○ User-defined range Start: End: (Coordinates: HIV1/SIV- HXB2, HIV2- MAC239)
Subtype: All
DNA/Protein: DNA
Format: Fasta

- **Web alignments**

What sequences are included

The alignments presented on the web differ from the ones that are printed in the compendia. The whole genome alignments are complete, meaning that they contain all complete genome sequences we have, including very similar ones.

The gene/protein alignments contain all complete gene sequences we have, with the important exception that very similar sequences (e.g. multiple clones from one isolate, multiple sequences from one person) have been deleted. The selection was made on the basis of phylogenetic trees: from tight clusters of sequences, one representative was retained and the others were removed from the alignment. An exception has been made for HXB2 and LAI, as these are important lab strains that are frequently used in experiments.

Alll(complete) = one per patient, all sequences for which we have a complete genome, or a complete gene.

Subtype Reference = 4 representatives of each subtype, plus one of each Circulating intersubtype recombinant form (CRF) of the M group, plus 4 O group, N group, P group and SIV-CPZ

Consensus/Ancestral computed from master alignment periodically.

HIV-2/SIV-SMM and primate lentivirus alignments also available here.

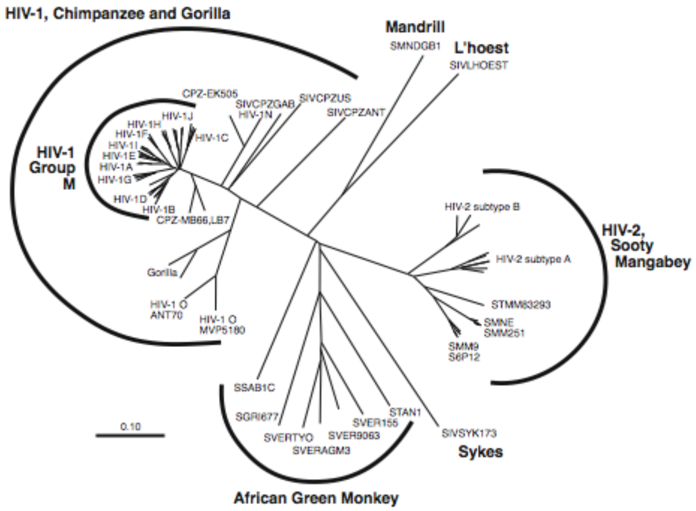
Los Alamos NATIONAL LABORATORY

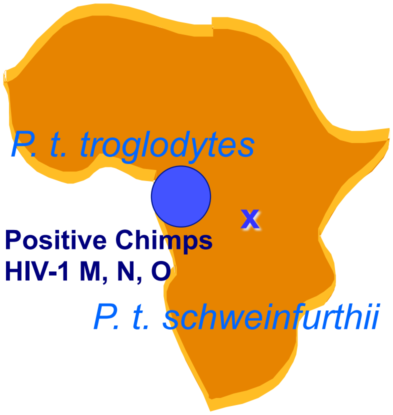
SIV/PLV Alignments

- Any non-human lentivirus is a SIV (or primate lentivirus), not just the SIV-SMM/SIV-MAC group from Sooty mangabeys.
- HIV-1s (M, N, O and P groups) are related to the SIV-CPZs from the chimps (*P. t. troglodytes)* and SIV-GORs from gorillas. We describe these alignments as HIV-1/CPZ.
- HIV-2s and SIV-MACs are related to SIV-SMMs from Sooty mangabeys. We describe these alignments as HIV-2/SMM.
- Dozens of other diverse non-human primates, such as African green monkeys, carry species-specific SIVs.
- Alignments of the diverse SIVs, plus HIVs, can help to identify highly conserved codons and other features. We describe these alignments as "other SIV" or HIV-1/HIV-2/SIV.

Primate Lentiviruses

Alignments: http://www.hiv.lanl.gov/content/hiv-db/ALIGN_CURRENT/ALIGN-INDEX.html



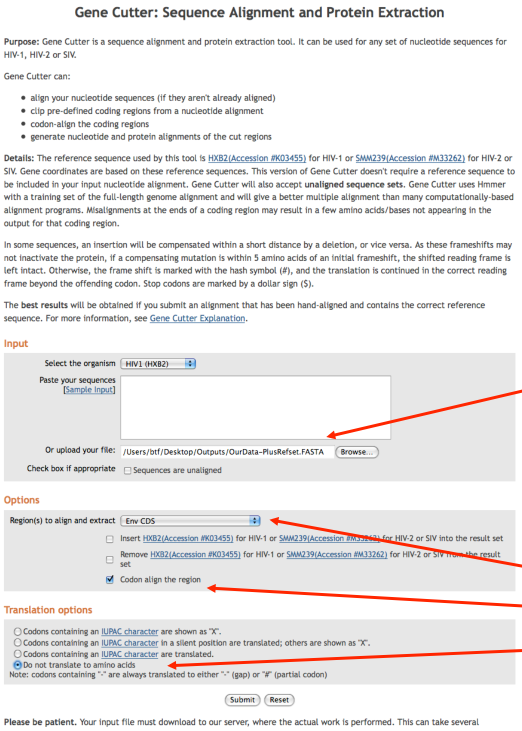


Van Heuverswyn, Nature 2006
Keele, *Science* 2006
Corbet, *J. Virol* 2000
Foley, HIV database

Gene Cutter

- Unconventional Alignment/Homology program
- “Cuts out” specified genes and proteins from sets of DNA sequences
  - □ Aligns to HXB2 via HMMer (or to SIV-Mac239 for HIV-2 and SIV-SMM)
  - □ Splits input sequences into genes, if desired
  - □ Aligns DNA sequences by codon, and translates them (including interpretation of IUPAC codes such as R for purine)
- Useful for processing new sequence data
  - □ annotating full length genomes
  - □ pulling out regions of interest from raw sequence data
- For each gene/region, maintains a list of anomalies
  - □ stop codons
  - □ codons containing multi-state characters
  - □ codons containing indels
- Input sequences may be aligned or unaligned
- Results may be better if the HXB2 sequence is included as a reference in your input file

GeneCutter



Gene Cutter: Sequence Alignment and Protein Extraction

Purpose: Gene Cutter is a sequence alignment and protein extraction tool. It can be used for any set of nucleotide sequences for HIV-1, HIV-2 or SIV.

Gene Cutter can:

- align your nucleotide sequences (if they aren't already aligned)
- clip pre-defined coding regions from a nucleotide alignment
- codon-align the coding regions
- generate nucleotide and protein alignments of the cut regions

Details: The reference sequence used by this tool is HXB2(Accession #K03455) for HIV-1 or SMM239(Accession #M33262) for HIV-2 or SIV. Gene coordinates are based on these reference sequences. This version of Gene Cutter doesn't require a reference sequence to be included in your input nucleotide alignment. Gene Cutter will also accept **unaligned sequence sets.** Gene Cutter uses Hmmer with a training set of the full-length genome alignment and will give a better multiple alignment than many computationally-based alignment programs. Misalignments at the ends of a coding region may result in a few amino acids/bases not appearing in the output for that coding region.

In some sequences, an insertion will be compensated within a short distance by a deletion, or vice versa. As these frameshifts may not inactivate the protein, if a compensating mutation is within 5 amino acids of an initial frameshift, the shifted reading frame is left intact. Otherwise, the frame shift is marked with the hash symbol (#), and the translation is continued in the correct reading frame beyond the offending codon. Stop codons are marked by a dollar sign ($).

The **best results** will be obtained if you submit an alignment that has been hand-aligned and contains the correct reference sequence. For more information, see Gene Cutter Explanation.

Input

Select the organism: HIV1 (HXB2)

Paste your sequences [Sample Input]

Or upload your file: /Users/btf/Desktop/Outputs/OurData-PlusRefset.FASTA Browse...

Check box if appropriate: □ Sequences are unaligned

Options

Region(s) to align and extract: Env CDS

□ Insert HXB2(Accession #K03455) for HIV-1 or SMM239(Accession #M33262) for HIV-2 or SIV into the result set

□ Remove HXB2(Accession #K03455) for HIV-1 or SMM239(Accession #M33262) for HIV-2 or SIV from the result set

☑ Codon align the region

Translation options

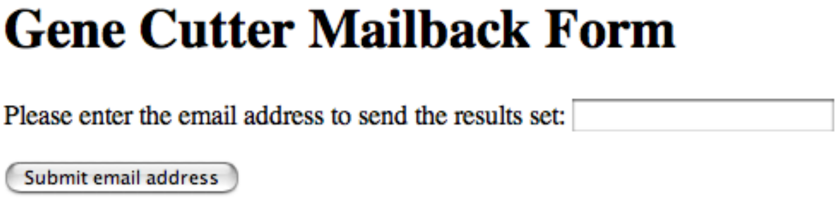
○ Codons containing an IUPAC character are shown as "X".
○ Codons containing an IUPAC character in a silent position are translated; others are shown as "X".
○ Codons containing an IUPAC character are translated.
● Do not translate to amino acids
Note: codons containing "-" are always translated to either "-" (gap) or "#" (partial codon)

Please be patient. Your input file must download to our server, where the actual work is performed. This can take several

Input is our data plus the “reference Set” and any other sequences we chose to add from the search interface.
Input: GeneCutterInput.FASTA
Output: GeneCutterOutputAll.FASTA

For this exercise, we want the Env gene, codon aligned, but not translated to proteins.
Output: GeneCutterOutputEnv.FASTA

GeneCutter Results



Gene Cutter Mailback Form

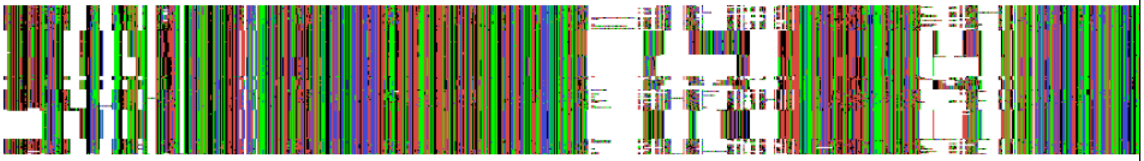
Please enter the email address to send the results set:

- Results are stored on our server
  - □ An HTML link is e-mailed to the user when the run is complete
  - □ For this workshop, we will provide example.

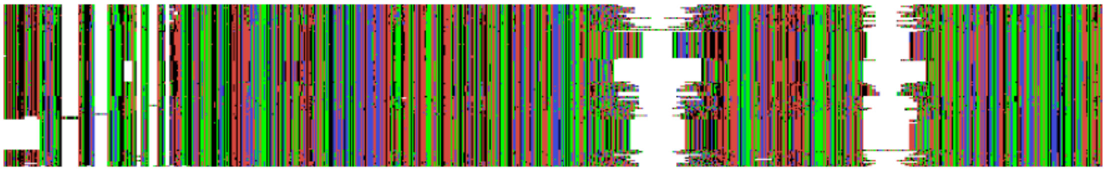
GeneCutter Result

Result saved in Outputs folder
Alignments viewed with Pixel
http://www.hiv.lanl.gov/content/sequence/pixel/pixel.html

Our data aligned to reference set by search tool:
GeneCutterInput.FASTA
(output of search and tree build was input to GeneCutter)



Our data aligned to reference set by GeneCutter:
Outputs: GeneCutterOutputENV.FASTA



Can also be viewed with BioEdit, Se-Al or other multiple sequence alignment editors.

Treemaker

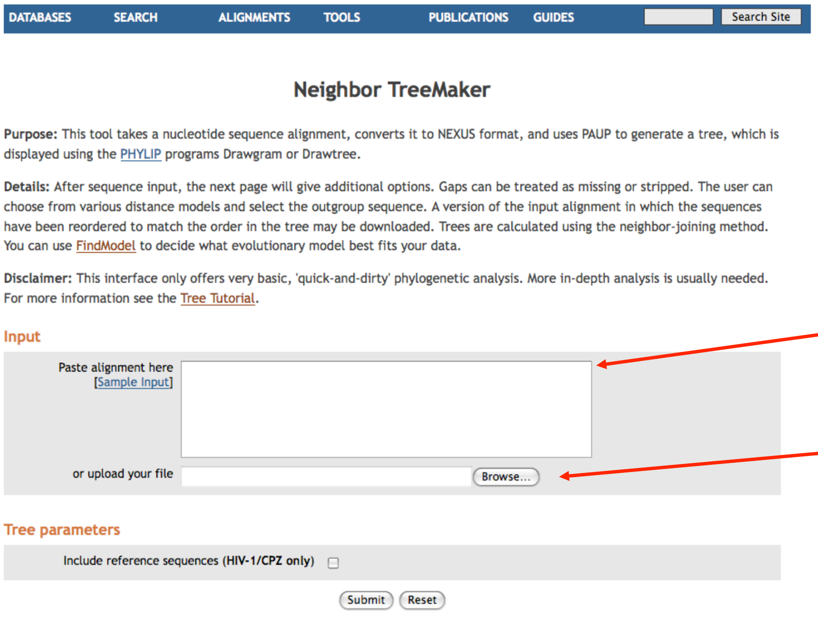
Check for phylogenetic relatives:

- TreeMaker produces a Neighbor Joining tree for a quick comparison
- TreeMaker uses PAUP* for its calculations; a few model options are available
- Reference sequences can be included, and are aligned to the input automatically
- Trees are displayed using PHYLIP and ATV
- The alignment used for the tree can also be downloaded
- A Phyml interface is also available

http://www.hiv.lanl.gov/content/sequence/PHYML/interface.html

Los Alamos
NATIONAL LABORATORY

http://www.hiv.lanl.gov/components/sequence/HIV/treemaker/treemaker.html



Neighbor TreeMaker

Purpose: This tool takes a nucleotide sequence alignment, converts it to NEXUS format, and uses PAUP to generate a tree, which is displayed using the PHYLIP programs Drawgram or Drawtree.

Details: After sequence input, the next page will give additional options. Gaps can be treated as missing or stripped. The user can choose from various distance models and select the outgroup sequence. A version of the input alignment in which the sequences have been reordered to match the order in the tree may be downloaded. Trees are calculated using the neighbor-joining method. You can use FindModel to decide what evolutionary model best fits your data.

Disclaimer: This interface only offers very basic, 'quick-and-dirty' phylogenetic analysis. More in-depth analysis is usually needed. For more information see the Tree Tutorial.

Input

Paste alignment here
[Sample Input]

or upload your file Browse...

Tree parameters

Include reference sequences (HIV-1/CPZ only) □

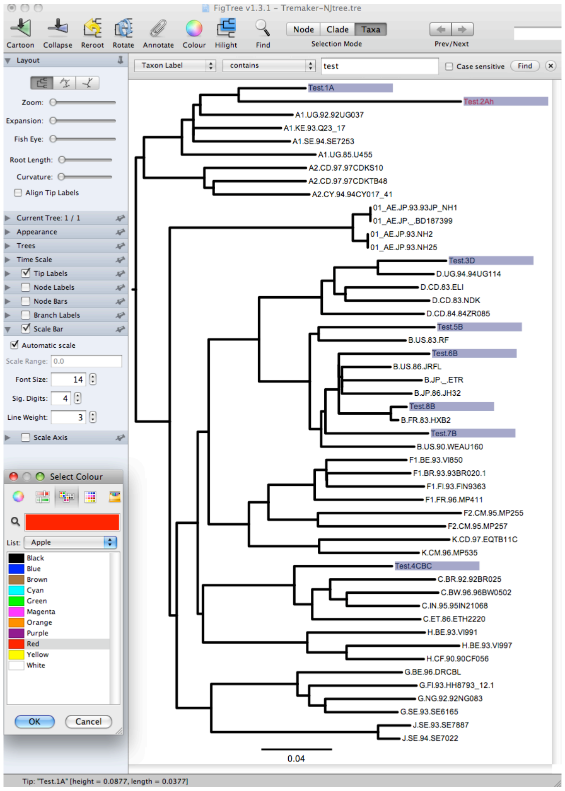
Submit Reset

Paste or type a DNA **alignment** here.

OR upload an alignment file here.

Los Alamos
NATIONAL LABORATORY

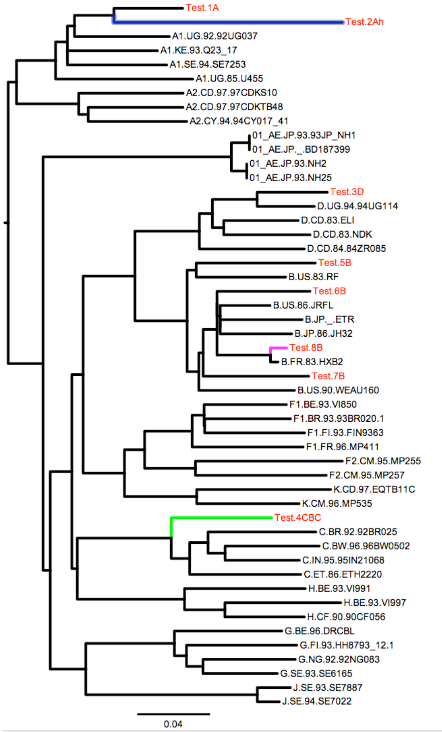
http://tree.bio.ed.ac.uk/software/figtree/



Taxon Label · contains · test · ☐ Case sensitive · Find

☑ Tip Labels · ☐ Node Labels · ☐ Node Bars · ☐ Branch Labels · ☑ Scale Bar · ☑ Automatic scale · Scale Range: 0.0 · Font Size: 14 · Sig. Digits: 4 · Line Weight: 3 · List: Apple

Tip: "Test.1A" [height = 0.0877, length = 0.0377]



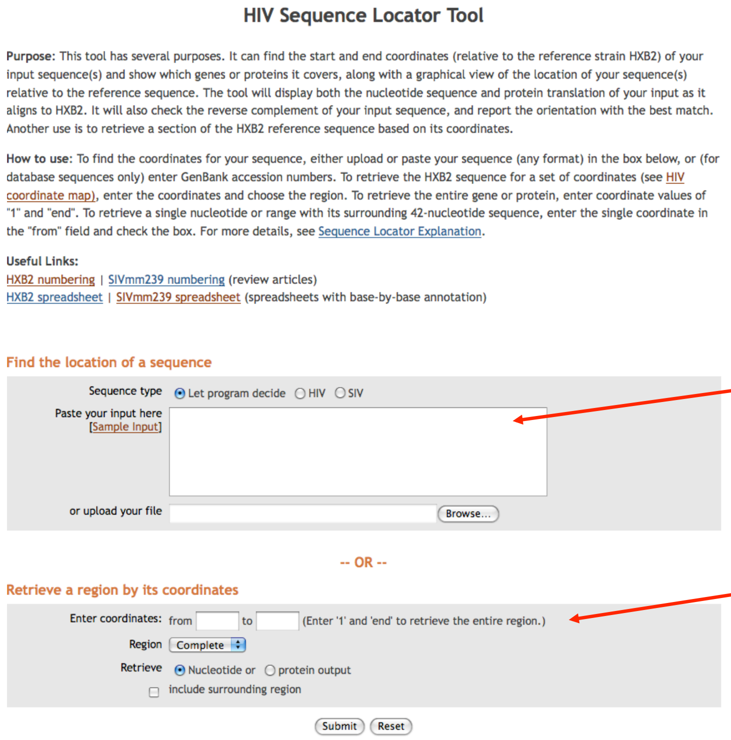
Los Alamos
NATIONAL LABORATORY

HIV/SIV Sequence Locator Tool

- Instantly computes position numbers of DNA or protein fragments relative to a reference strain (HXB2r for HIV-1, SMM239 for SIV)
  - Such numbers, often included in the literature, are frequently incorrect
- Shows the location of the sequence on an HIV map
- Presents protein translations of DNA sequences
- Can be used for input into the search interface, to align a new sequence you have generated with the database set
- Can also retrieve reference sequences
  - by coordinates (range of base or amino-acid positions)
  - by single position (retrieves flanking sequences)

Los Alamos
NATIONAL LABORATORY

http://www.hiv.lanl.gov/content/sequence/LOCATE/locate.html



HIV Sequence Locator Tool

Purpose: This tool has several purposes. It can find the start and end coordinates (relative to the reference strain HXB2) of your input sequence(s) and show which genes or proteins it covers, along with a graphical view of the location of your sequence(s) relative to the reference sequence. The tool will display both the nucleotide sequence and protein translation of your input as it aligns to HXB2. It will also check the reverse complement of your input sequence, and report the orientation with the best match. Another use is to retrieve a section of the HXB2 reference sequence based on its coordinates.

How to use: To find the coordinates for your sequence, either upload or paste your sequence (any format) in the box below, or (for database sequences only) enter GenBank accession numbers. To retrieve the HXB2 sequence for a set of coordinates (see HIV coordinate map), enter the coordinates and choose the region. To retrieve the entire gene or protein, enter coordinate values of "1" and "end". To retrieve a single nucleotide or range with its surrounding 42-nucleotide sequence, enter the single coordinate in the "from" field and check the box. For more details, see Sequence Locator Explanation.

Useful Links:
HXB2 numbering | SIVmm239 numbering (review articles)
HXB2 spreadsheet | SIVmm239 spreadsheet (spreadsheets with base-by-base annotation)

Find the location of a sequence

Sequence type ● Let program decide ○ HIV ○ SIV

Paste your input here [Sample Input]

or upload your file Browse...

-- OR --

Retrieve a region by its coordinates

Enter coordinates: from to (Enter '1' and 'end' to retrieve the entire region.)

Region Complete

Retrieve ● Nucleotide or ○ protein output

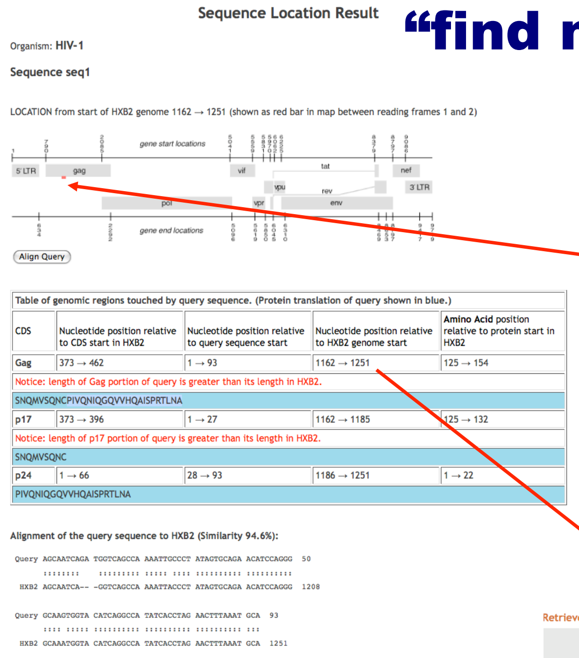
□ include surrounding region

Submit Reset

Paste or type a DNA or protein sequence here.

OR enter numeric coordiantes here.

Sequence Locator: "find my sequence"



Sequence Location Result

Organism: **HIV-1**

Sequence seq1

LOCATION from start of HXB2 genome 1162 → 1251 (shown as red bar in map between reading frames 1 and 2)

Table of genomic regions touched by query sequence. (Protein translation of query shown in blue.)

CDS	Nucleotide position relative to CDS start in HXB2	Nucleotide position relative to query sequence start	Nucleotide position relative to HXB2 genome start	**Amino Acid** position relative to protein start in HXB2
Gag	373 → 462	1 → 93	1162 → 1251	125 → 154
Notice: length of Gag portion of query is greater than its length in HXB2.				
SNQMVSQNCPIVQNIQGQVVHQAISPRTLNA				
p17	373 → 396	1 → 27	1162 → 1185	125 → 132
Notice: length of p17 portion of query is greater than its length in HXB2.				
SNQMVSQNC				
p24	1 → 66	28 → 93	1186 → 1251	1 → 22
PIVQNIQGQVVHQAISPRTLNA				

Alignment of the query sequence to HXB2 (Similarity 94.6%):

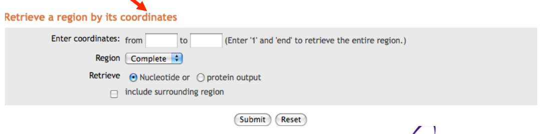
```
Query AGCAATCAGA TGGTCAGCCA AAATTGCCCT ATAGTGCAGA ACATCCAGGG  50
      ::::::::    ::::::::: ::::: :::: :::::::::: ::::::::::
 HXB2 AGCAATCA-- -GGTCAGCCA AAATTACCCT ATAGTGCAGA ACATCCAGGG  1208

Query GCAAGTGGTA CATCAGGCCA TATCACCTAG AACTTTAAAT GCA  93
      :::: ::::: :::::::::: :::::::::: :::::::::: :::
 HXB2 GCAAATGGTA CATCAGGCCA TATCACCTAG AACTTTAAAT GCA  1251
```

Result for Sample Input DNA Query sequence

Location in genome mapped in red.

Numeric coordinates useful for entry on search form

DNA and protein sequence displayed



Retrieve a region by its coordinates

Enter coordinates: from to (Enter '1' and 'end' to retrieve the entire region.)

Region Complete

Retrieve ● Nucleotide or ○ protein output

□ include surrounding region

Submit Reset

Sequence Locator:
“Retrieve from coordinates”

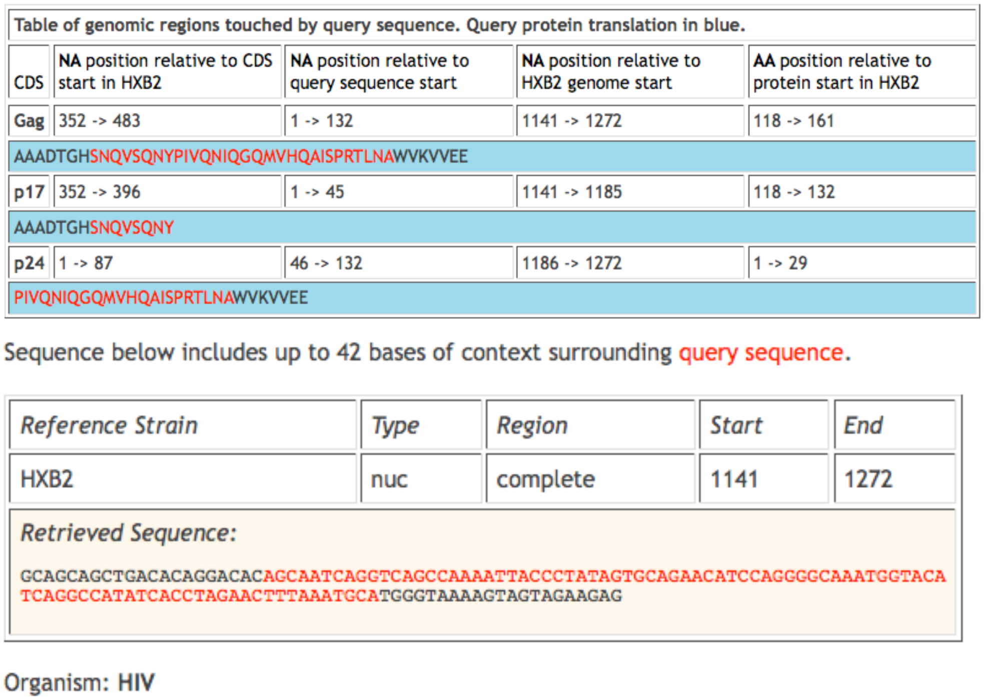


Table of genomic regions touched by query sequence. Query protein translation in blue.				
CDS	NA position relative to CDS start in HXB2	NA position relative to query sequence start	NA position relative to HXB2 genome start	AA position relative to protein start in HXB2
Gag	352 -> 483	1 -> 132	1141 -> 1272	118 -> 161
AAADTGHSNQVSQNYPIVQNIQGQMVHQAISPRTLNAWVKVVEE				
p17	352 -> 396	1 -> 45	1141 -> 1185	118 -> 132
AAADTGHSNQVSQNY				
p24	1 -> 87	46 -> 132	1186 -> 1272	1 -> 29
PIVQNIQGQMVHQAISPRTLNAWVKVVEE				

Sequence below includes up to 42 bases of context surrounding query sequence.

Reference Strain	*Type*	*Region*	*Start*	*End*
HXB2	nuc	complete	1141	1272

Retrieved Sequence:

```
GCAGCAGCTGACACAGGACACAGCAATCAGGTCAGCCAAAATTACCCTATAGTGCAGAACATCCAGGGGCAAATGGTACA
TCAGGCCATATCACCTAGAACTTTAAATGCATGGGTAAAAGTAGTAGAAGAG
```

Organism: HIV

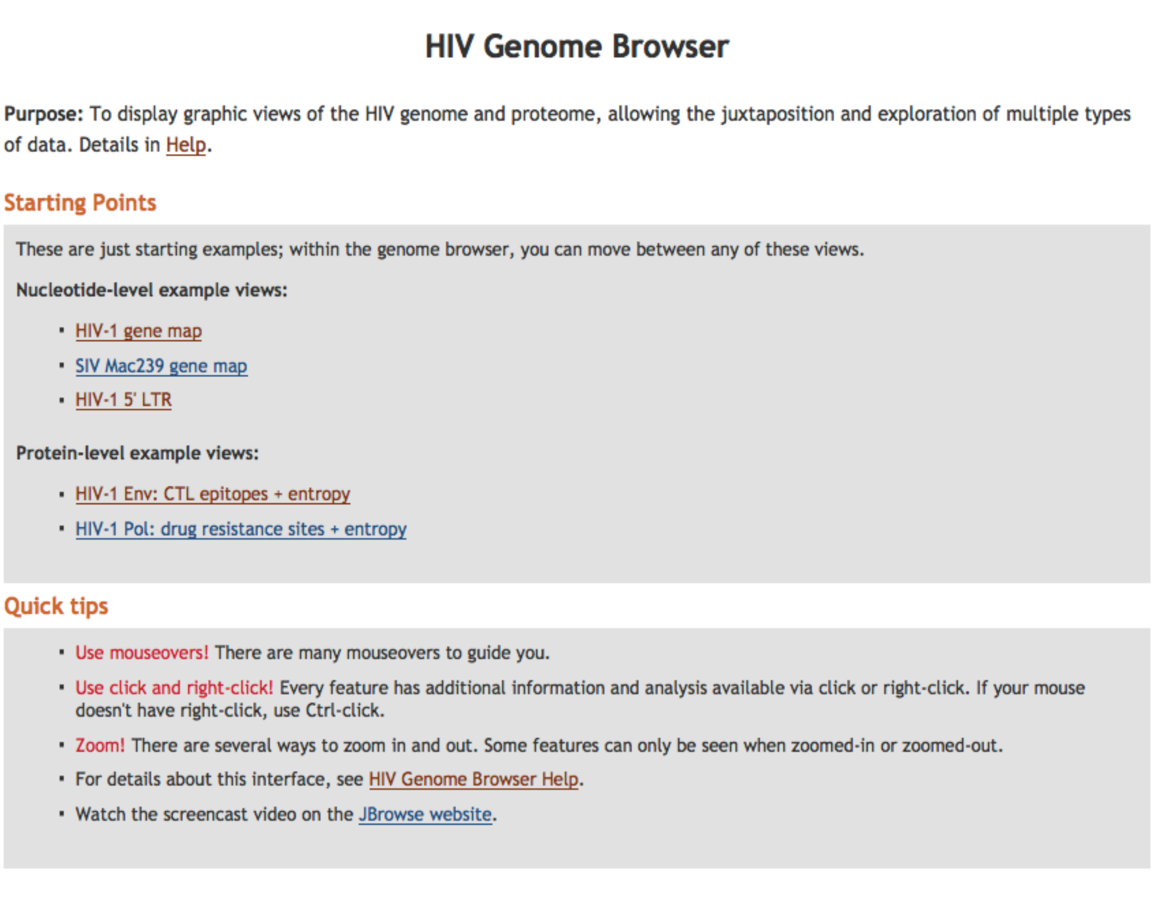
HIV Genome Browser:

- Dreamed of by Christian Brander and designed by Shihai Feng, with the help from Jennifer Macke, Brian Foley, Jim Szinger, Karina Yusim

- A customization of Jbrowse Genome Browser, built to incorporate many sources of information from the LANL HIV Sequence and Immunology databases.

- A one-stop source of information about HIV genome and immunological data. It retrieves the vast and diverse information available at HIV Immunology database and allows to look at the whole HIV genome as well as zoom in to a region of interest and see all information we have in the database about this region
  - HXB2 gene map, HXB2 sub-protein map, Mac239 map
  - Overlapping epitopes, antibody binding sites
  - HXB2 coding sites of interest (e.g. functional domains, drug resistance sites, motifs, glycosylation sites, etc.)
  - HXB2 LTR sites of interest (RNA structural elements, primer binding sites, etc.)
  - Neutralizing Ab contact residues, signatures and other NAb-associated features
  - HIV sequence variability (Entropy: M group, B clade, C clade)
  - Links to the database annotation, alignments, tools, Pubmed etc.

- DNA- and Protein-level views are available



HIV Genome Browser

Purpose: To display graphic views of the HIV genome and proteome, allowing the juxtaposition and exploration of multiple types of data. Details in Help.

Starting Points

These are just starting examples; within the genome browser, you can move between any of these views.

Nucleotide-level example views:

- HIV-1 gene map
- SIV Mac239 gene map
- HIV-1 5' LTR

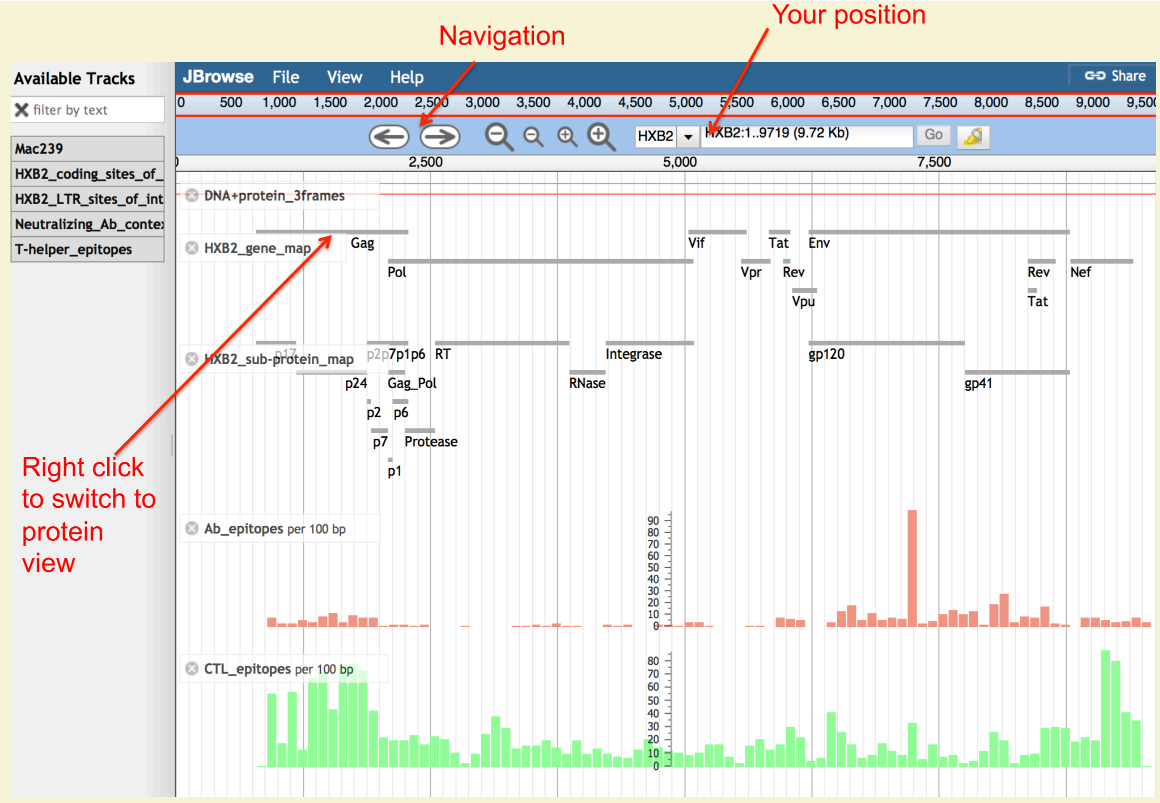
Protein-level example views:

- HIV-1 Env: CTL epitopes + entropy
- HIV-1 Pol: drug resistance sites + entropy

Quick tips

- Use mouseovers! There are many mouseovers to guide you.
- Use click and right-click! Every feature has additional information and analysis available via click or right-click. If your mouse doesn't have right-click, use Ctrl-click.
- Zoom! There are several ways to zoom in and out. Some features can only be seen when zoomed-in or zoomed-out.
- For details about this interface, see HIV Genome Browser Help.
- Watch the screencast video on the JBrowse website.

HIV Genome Browser: Nucleotide view

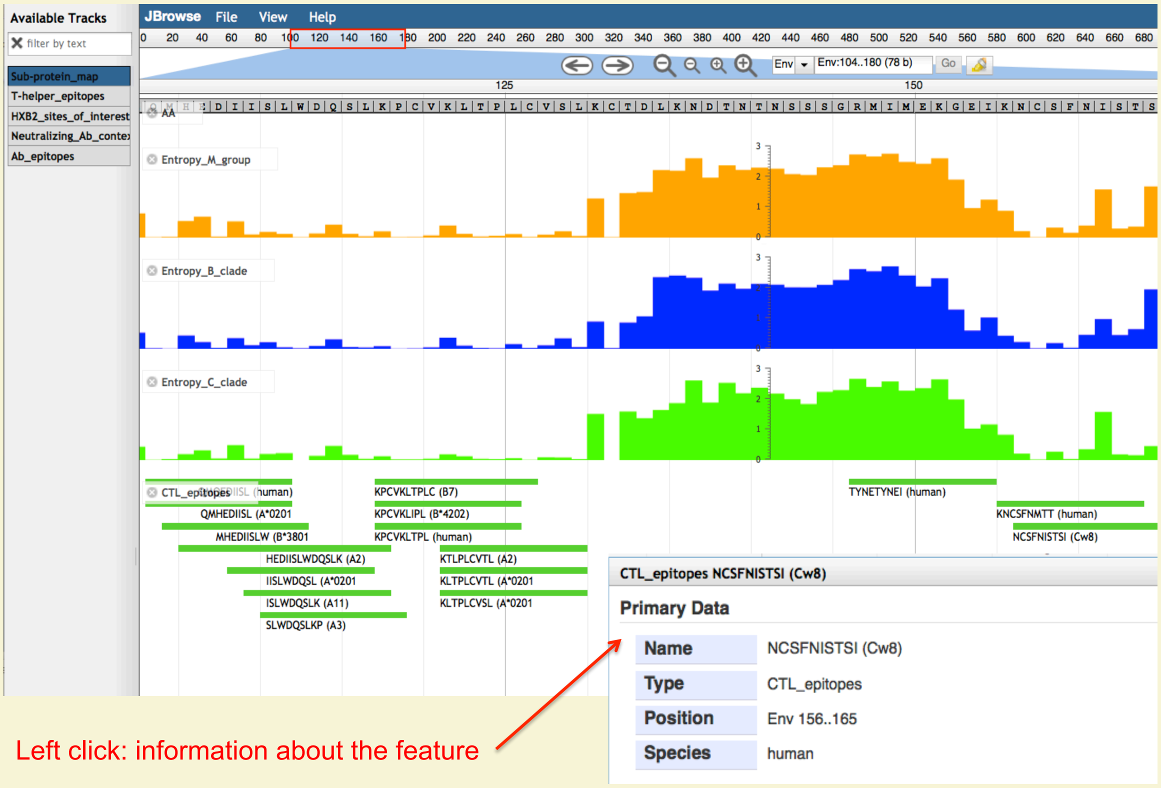


Navigation

Your position

Right click to switch to protein view

HIV Genome Browser: Protein view

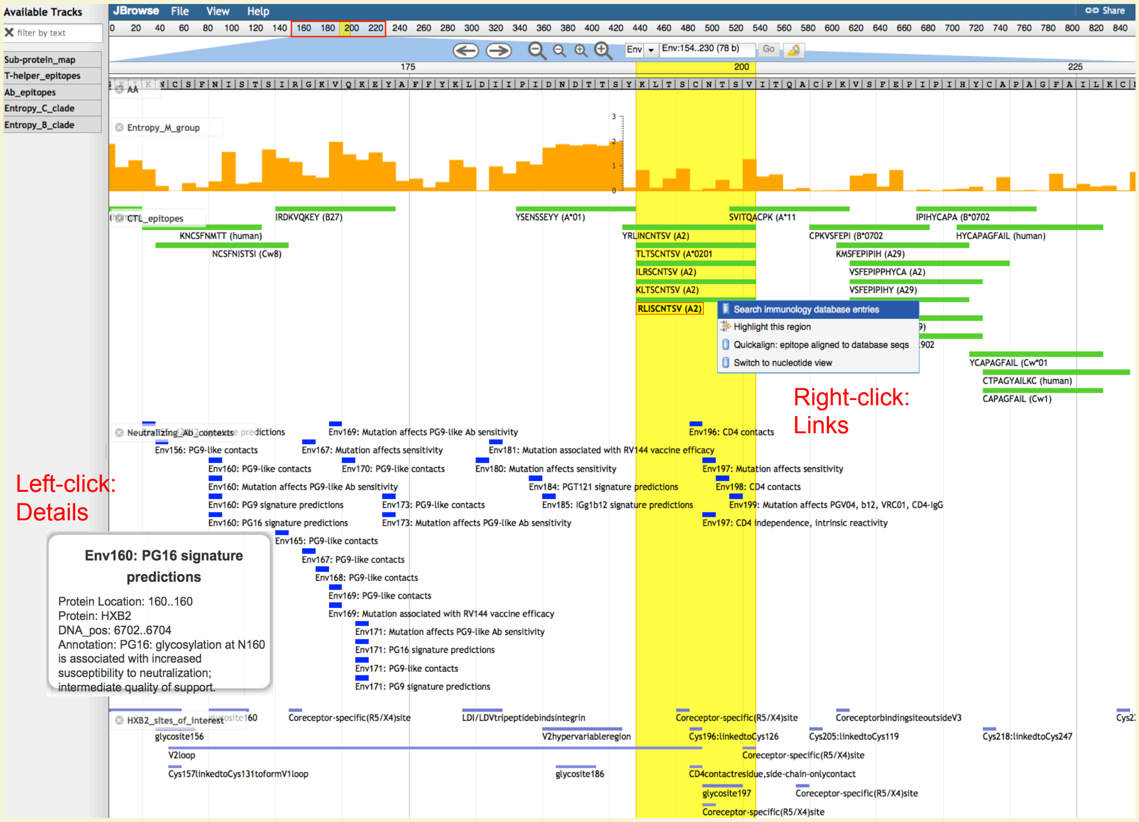


CTL_epitopes NCSFNISTSI (Cw8)

Primary Data

Name	NCSFNISTSI (Cw8)
Type	CTL_epitopes
Position	Env 156..165
Species	human

Left click: information about the feature


Available Tracks
filter by text
Sub-protein_map
T-helper_epitopes
Ab_epitopes
Entropy_C_clade
Entropy_B_clade
JBrowse
File
View
Help
Share
Env
Env:154..230 (78 b)
Go
175
200
225
AA
Entropy_M_group
CTL_epitopes
KNCSFNMTT (human)
NCSFNISTSI (Cw8)
IRDKVQKEY (B27)
YSENSSEYY (A*01)
YRLINCNTSV (A2)
TLTSCNTSV (A*0201
ILRSCNTSV (A2)
KLTSCNTSV (A2)
RLISCNTSV (A2)
SVITQACPK (A*11
CPKVSFEPI (B*0702
KMSFEPIPIH (A29)
VSFEPIPPHYCA (A2)
VSFEPIPIHY (A29)
IPIHYCAPA (B*0702
HYCAPAGFAIL (human)
YCAPAGFAIL (Cw*01
CTPAGYAILKC (human)
CAPAGFAIL (Cw1)
Search immunology database entries
Highlight this region
Quickalign: epitope aligned to database seqs
Switch to nucleotide view
Neutralizing_Ab_contexts
Env156: PG9-like contacts
Env160: PG9-like contacts
Env160: Mutation affects PG9-like Ab sensitivity
Env160: PG9 signature predictions
Env160: PG16 signature predictions
Env165: PG9-like contacts
Env167: Mutation affects sensitivity
Env167: PG9-like contacts
Env168: PG9-like contacts
Env169: Mutation affects PG9-like Ab sensitivity
Env169: PG9-like contacts
Env169: Mutation associated with RV144 vaccine efficacy
Env170: PG9-like contacts
Env171: Mutation affects PG9-like Ab sensitivity
Env171: PG16 signature predictions
Env171: PG9-like contacts
Env171: PG9 signature predictions
Env173: PG9-like contacts
Env173: Mutation affects PG9-like Ab sensitivity
Env180: Mutation affects sensitivity
Env181: Mutation associated with RV144 vaccine efficacy
Env184: PGT121 signature predictions
Env185: IGg1b12 signature predictions
Env196: CD4 contacts
Env197: Mutation affects sensitivity
Env198: CD4 contacts
Env199: Mutation affects PGV04, b12, VRC01, CD4-IgG
Env197: CD4 independence, intrinsic reactivity
HXB2_sites_of_interest
glycosite160
glycosite156
V2loop
Cys157linkedtoCys131toformV1loop
Coreceptor-specific(R5/X4)site
LDI/LDVtripeptidebindsintegrin
V2hypervariableregion
glycosite186
Cys196:linkedtoCys126
Cys205:linkedtoCys119
CoreceptorbindingsiteoutsideV3
Cys218:linkedtoCys247
CD4contactresidue,side-chain-onlycontact
glycosite197

Right-click: Links

Left-click: Details

Env160: PG16 signature predictions

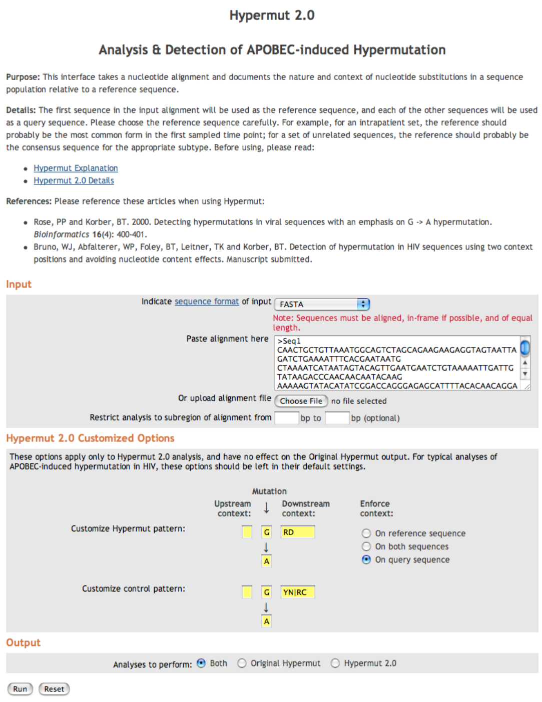
Protein Location: 160..160
Protein: HXB2
DNA_pos: 6702..6704
Annotation: PG16: glycosylation at N160 is associated with increased susceptibility to neutralization; intermediate quality of support.

HIV genome browser: more possibilities ?

- Data of how heavily sequenced each genome region is (we are getting questions sometimes why some regions don't return a lot of sequences on the sequence search interface)
- Show subtype consensus sequences
- CTL Epitope variants (we currently have a database of ~3000 CTL epitope variant records and started Helper epitope variants)
- Categorize heavily loaded tracks. For example, provide separate tracks for Drug resistance, CD4 contact residues, Ab contact residues, Glycosylation sites etc
- Links to structure
- Suggestions ?

Los Alamos
NATIONAL LABORATORY

Hypermutation



Hypermut 2.0

Analysis & Detection of APOBEC-induced Hypermutation

Purpose: This interface takes a nucleotide alignment and documents the nature and context of nucleotide substitutions in a sequence population relative to a reference sequence.

Details: The first sequence in the input alignment will be used as the reference sequence, and each of the other sequences will be used as a query sequence. Please choose the reference sequence carefully. For example, for an intrapatient set, the reference should probably be the most common form in the first sampled time point; for a set of unrelated sequences, the reference should probably be the consensus sequence for the appropriate subtype. Before using, please read:

- Hypermut Explanation
- Hypermut 2.0 Details

References: Please reference these articles when using Hypermut:

- Rose, PP and Korber, BT. 2000. Detecting hypermutations in viral sequences with an emphasis on G -> A hypermutation. *Bioinformatics* **16**(4): 400-401.
- Bruno, WJ, Abfalterer, WP, Foley, BT, Leitner, TK and Korber, BT. Detection of hypermutation in HIV sequences using two context positions and avoiding nucleotide content effects. Manuscript submitted.

Input

Indicate sequence format of input: FASTA

Note: Sequences must be aligned, in-frame if possible, and of equal length.

Paste alignment here:

```
>Seq1
CAACTGCTGTTAAATGGCAGTCTAGCAGAAGAAGAGGTAGTAATTA
GATCTGAAAATTTCACGAATAATG
CTAAAATCATAATAGTACAGTTGAATGAATCTGTAAAAATTGATTG
TATAAGACCCAACAACAATACAAG
AAAAAGTATACATATCGGACCAGGGAGAGCATTTTACACAACAGGA
```

Or upload alignment file: Choose File no file selected

Restrict analysis to subregion of alignment from ____ bp to ____ bp (optional)

Hypermut 2.0 Customized Options

These options apply only to Hypermut 2.0 analysis, and have no effect on the Original Hypermut output. For typical analyses of APOBEC-induced hypermutation in HIV, these options should be left in their default settings.

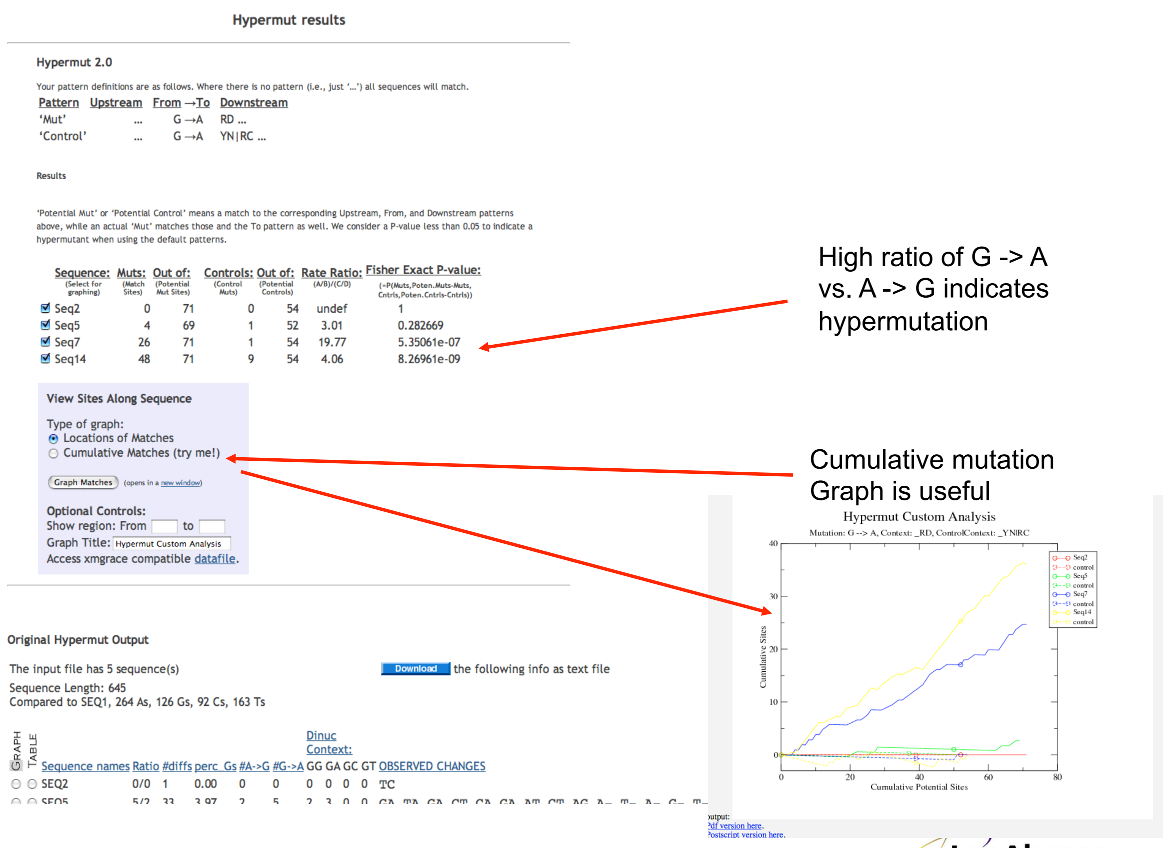
	Upstream context:	Mutation ↓	Downstream context:	Enforce context:
Customize Hypermut pattern:		G ↓ A	RD	○ On reference sequence ○ On both sequences ● On query sequence
Customize control pattern:		G ↓ A	YN|RC	

Output

Analyses to perform: ● Both ○ Original Hypermut ○ Hypermut 2.0

- Detects APOBEC related A->G hypermutation as default
- Can be adapted to detect any fuzzy motif in relation to a control pattern

Los Alamos
NATIONAL LABORATORY



Hypermut results

Hypermut 2.0

Your pattern definitions are as follows. Where there is no pattern (i.e., just '...') all sequences will match.

Pattern	Upstream	From →To	Downstream
'Mut'	...	G →A	RD ...
'Control'	...	G →A	YN|RC ...

Results

'Potential Mut' or 'Potential Control' means a match to the corresponding Upstream, From, and Downstream patterns above, while an actual 'Mut' matches those and the To pattern as well. We consider a P-value less than 0.05 to indicate a hypermutant when using the default patterns.

Sequence: (Select for graphing)	Muts: (Match Sites)	Out of: (Potential Mut Sites)	Controls: (Control Muts)	Out of: (Potential Controls)	Rate Ratio: (A/B)/(C/D)	Fisher Exact P-value: (=P(Muts,Poten.Muts-Muts, Cntrls,Poten.Cntrls-Cntrls))
☑ Seq2	0	71	0	54	undef	1
☑ Seq5	4	69	1	52	3.01	0.282669
☑ Seq7	26	71	1	54	19.77	5.35061e-07
☑ Seq14	48	71	9	54	4.06	8.26961e-09

View Sites Along Sequence

Type of graph:
● Locations of Matches
○ Cumulative Matches (try me!)

Graph Matches (opens in a new window)

Optional Controls:
Show region: From □ to □
Graph Title: Hypermut Custom Analysis
Access xmgrace compatible datafile.

Original Hypermut Output

The input file has 5 sequence(s) Download the following info as text file

Sequence Length: 645
Compared to SEQ1, 264 As, 126 Gs, 92 Cs, 163 Ts

GRAPH	TABLE	Sequence names	Ratio	#diffs	perc_Gs	#A->G	#G->A	Dinuc Context: GG	GA	GC	GT	OBSERVED CHANGES
○	○	SEQ2	0/0	1	0.00	0	0	0	0	0	0	TC
○	○	SEQ5	5/2	33	3.97	2	5	2	3	0	0	GA TA GA CT GA GA AT CT AG A- T- A- G- T-

High ratio of G -> A vs. A -> G indicates hypermutation

Cumulative mutation Graph is useful

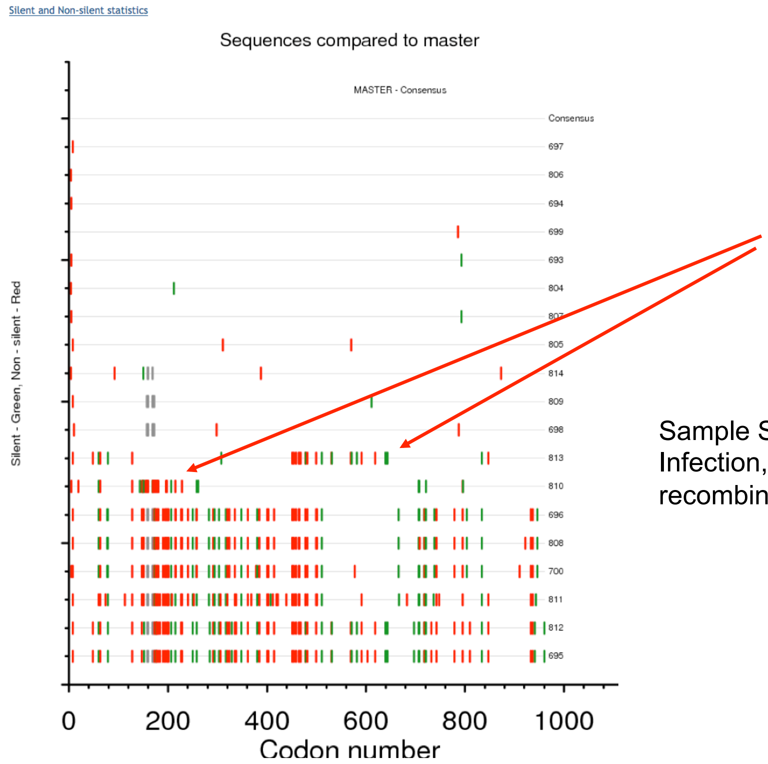
Hypermut Custom Analysis

Mutation: G --> A, Context: _RD, ControlContext: _YN|RC

output:
Pdf version here.
Postscript version here.

Highlighter

- Highlights mutations relative to a reference strain, particularly useful for intra-patient analyses.
- Highlights:
  - syn/non-syn
  - transition/transversion
  - Apobec motifs
- Sorts on similarity
- Visualize recombination of closely related sequences



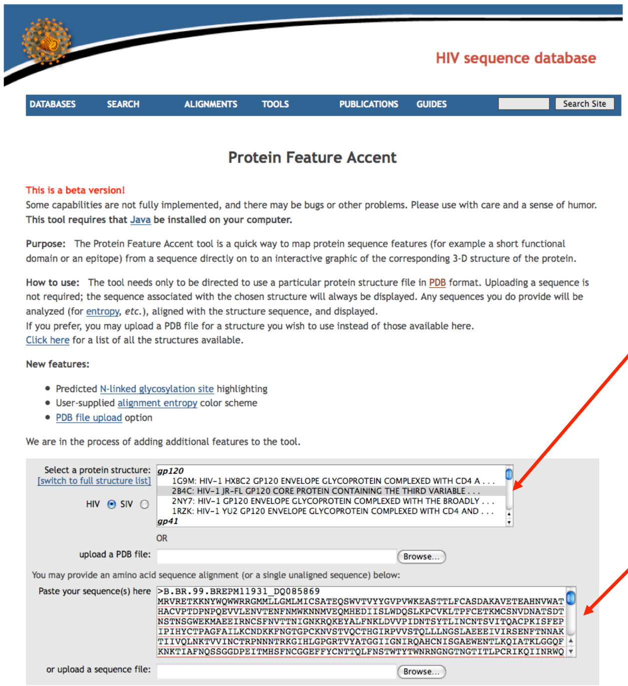
Nonrandom distribution of mutations evident.

Sample Set is from a possible dual Infection, with intra-subtype recombinants evident.

Protein Feature Accent

- Highlights region of interest in an HIV structure
- You can upload a PDB structure, or use one of our annotated Env structures
- You can upload your own alignment and get an entropy map

http://www.hiv.lanl.gov/content/sequence/PROTVIS/html/protvis.html



Protein Feature Accent

This is a beta version!
Some capabilities are not fully implemented, and there may be bugs or other problems. Please use with care and a sense of humor.
This tool requires that Java be installed on your computer.

Purpose: The Protein Feature Accent tool is a quick way to map protein sequence features (for example a short functional domain or an epitope) from a sequence directly on to an interactive graphic of the corresponding 3-D structure of the protein.

How to use: The tool needs only to be directed to use a particular protein structure file in PDB format. Uploading a sequence is not required; the sequence associated with the chosen structure will always be displayed. Any sequences you do provide will be analyzed (for entropy, *etc.*), aligned with the structure sequence, and displayed.
If you prefer, you may upload a PDB file for a structure you wish to use instead of those available here.
Click here for a list of all the structures available.

New features:

- Predicted N-linked glycosylation site highlighting
- User-supplied alignment entropy color scheme
- PDB file upload option

We are in the process of adding additional features to the tool.

Select a protein structure: [switch to full structure list]

gp120
1G9M: HIV–1 HXBC2 GP120 ENVELOPE GLYCOPROTEIN COMPLEXED WITH CD4 A . . .
2B4C: HIV–1 JR–FL GP120 CORE PROTEIN CONTAINING THE THIRD VARIABLE . . .
2NY7: HIV–1 GP120 ENVELOPE GLYCOPROTEIN COMPLEXED WITH THE BROADLY . . .
1RZK: HIV–1 YU2 GP120 ENVELOPE GLYCOPROTEIN COMPLEXED WITH CD4 AND . . .
gp41

HIV ● SIV ○

OR

upload a PDB file: Browse...

You may provide an amino acid sequence alignment (or a single unaligned sequence) below:

Paste your sequence(s) here

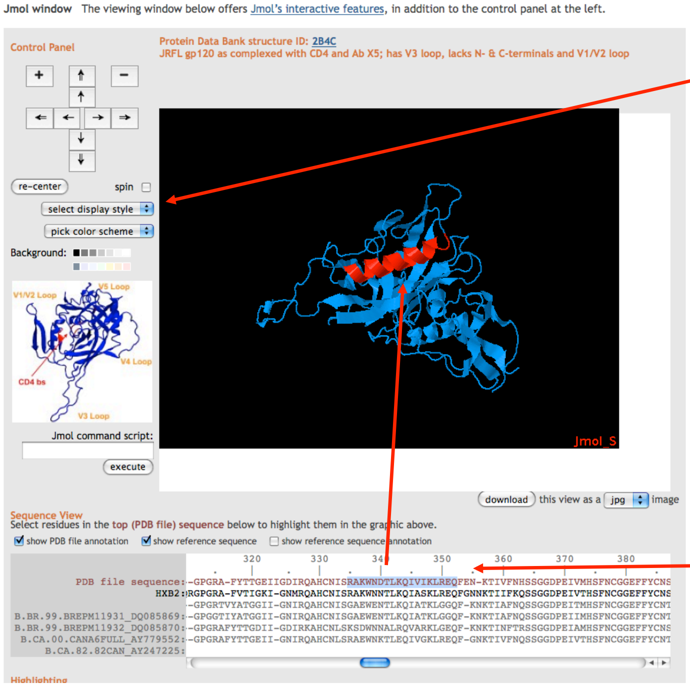
```
>B.BR.99.BREPM11931_DQ085869
MRVRETKKNYWQWWRRGMMLLGMLMICSATEQSWVTVYYGVPVWKEASTTLFCASDAKAVETEAHNVWAT
HACVPTDPNPQEVVLENVTENFNMWKNNMVEQMHEDIISLWDQSLKPCVKLTPFCETKMCSNVDNATSDT
NSTNSGWEKMAEEIRNCSFNVTTNIGNKRQKEYALFNKLDVVPIDNTSYTLINCNTSVITQACPKISFEP
IPIHYCTPAGFAILKCNDKKFNGTGPCKNVSTVQCTHGIRPVVSTQLLLNGSLAEEEIVIRSENFTNNAK
TIIVQLNKTVVINCTRPNNNTRKGIHLGPGRTVYATGGIIGNIRQAHCNISGAEWENTLKQIATKLGGQF
KNKTIAFNQSSGGDPEITMHSFNCGGEFFYCNTTQLFNSTWTYTWNRNGNGTNGTITLPCRIKQIINRWQ
```

or upload a sequence file: Browse...

List of “recommended” PDB entries

Only a gp120 alignment is provided so far. We hope to add others. You can paste in your own.

http://www.hiv.lanl.gov/content/sequence/PROTVIS/html/protvis.html



Jmol window The viewing window below offers Jmol’s interactive features, in addition to the control panel at the left.

Control Panel

Protein Data Bank structure ID: 2B4C
JRFL gp120 as complexed with CD4 and Ab X5; has V3 loop, lacks N- & C-terminals and V1/V2 loop

Jmol command script:

download this view as a jpg image

Sequence View
Select residues in the top (PDB file) sequence below to highlight them in the graphic above.

☑ show PDB file annotation ☑ show reference sequence ☐ show reference sequence annotation

```
                                  320       330       340       350       360       370       380
                                .    |    .    |    .    |    .    |    .    |    .    |    .    |    .
             PDB file sequence:-GPGRA-FYTTGEIIGDIRQAHCNISRAKWNDTLKQIVIKLREQFEN-KTIVFNHSSGGDPEIVMHSFNCGGEFFYCNS
                          HXB2:RGPGRA-FVTIGKI-GNMRQAHCNISRAKWNNTLKQIASKLREQFGNNKTIIFKQSSGGDPEIVTHSFNCGGEFFYCNS
                               -GPGRTVYATGGII-GNIRQAHCNISGAEWENTLKQIATKLGGQF-KNKTIAFNQSSGGDPEITMHSFNCGGEFFYCNT
B.BR.99.BREPM11931_DQ085869:-GPGGTIYATGGII-GNIRQAHCNISGAEWENTLKQIATKLGGQF-KNKTIAFNQSSGGDPEIIMHSFNCGGEFFYCNT
B.BR.99.BREPM11932_DQ085870:-GPGRAFYTTGDII-GDIRKAHCNLSKSDWNNALRQVARKLGEQF-KNKTINFTRSSGGDPEIAMHSFNCGGEFFYCNS
 B.CA.00.CANA6FULL_AY779552:-GPGRAFYTTGEII-GNIRQAHCNLSRAEWNKTLEQIVGKLREQF-GNKTIVFNQSSGGDPEIVTHSFNCGGEFFYCNT
      B.CA.82.82CAN_AY247225:
```

Highlighting

Many display options in JMol are “built in” to this web tool.
Use the JMol command script box below for other commands.

One of the color schemes is “color by entropy” based on divrsity in the alingment added below.

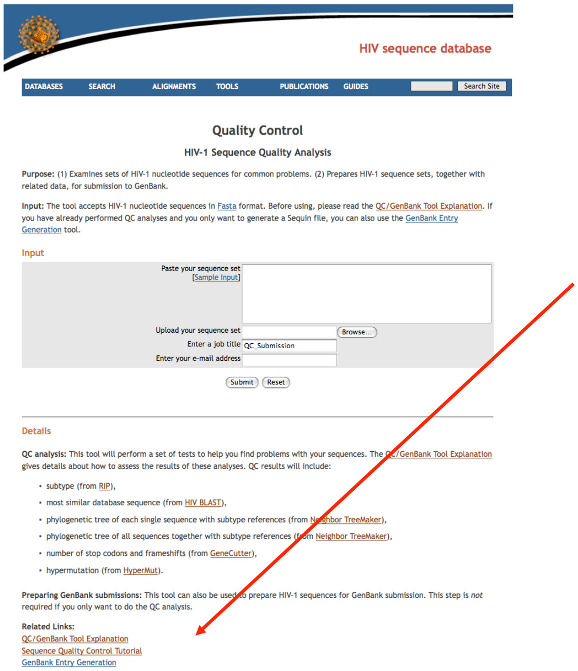
Selected region gets highlighted in structure

Quality Control Tool

- Built from existing HIV database tools
- GeneCutter
- RIP
- Hypermut
- Neighbor-joining Trees
- Output is an email containing a link to a summary report
- http://www.hiv.lanl.gov/content/hiv-db/QC/index.html (beta version)

Quality Control Tool

http://www.hiv.lanl.gov/content/sequence/QC/index



Quality Control

HIV-1 Sequence Quality Analysis

Purpose: (1) Examines sets of HIV-1 nucleotide sequences for common problems. (2) Prepares HIV-1 sequence sets, together with related data, for submission to GenBank.

Input: The tool accepts HIV-1 nucleotide sequences in Fasta format. Before using, please read the QC/GenBank Tool Explanation. If you have already performed QC analyses and you only want to generate a Sequin file, you can also use the GenBank Entry Generation tool.

Input

Paste your sequence set [Sample Input]

Upload your sequence set Browse...

Enter a job title QC_Submission

Enter your e-mail address

Submit Reset

Details

QC analysis: This tool will perform a set of tests to help you find problems with your sequences. The QC/GenBank Tool Explanation gives details about how to assess the results of these analyses. QC results will include:

- subtype (from RIP),
- most similar database sequence (from HIV BLAST),
- phylogenetic tree of each single sequence with subtype references (from Neighbor TreeMaker),
- phylogenetic tree of all sequences together with subtype references (from Neighbor TreeMaker),
- number of stop codons and frameshifts (from GeneCutter),
- hypermutation (from HyperMut).

Preparing GenBank submissions: This tool can also be used to prepare HIV-1 sequences for GenBank submission. This step is *not* required if you only want to do the QC analysis.

Related Links:
QC/GenBank Tool Explanation
Sequence Quality Control Tutorial
GenBank Entry Generation

Recently added shortcuts to GenBank entry creation tool.

Requires FASTA format sequences, and a comma separated values (CSV) file of annotations, as described on the help page.

http://www.hiv.lanl.gov/content/sequence/QC/field_help.html

Easy to enter in spreadsheet like EXCEL, and then export as CSV format.

Quality Control Tool

- Summary of results from analysis programs
- Click on each result to obtain full analysis
- Useful for helping to determine subtype, hypermutation, mislabeling of samples

Los Alamos
NATIONAL LABORATORY

Please leave any comments or suggestions with us:

Los Alamos
NATIONAL LABORATORY